

Applied Analysis

P. Ouwehand

Department of Mathematical Sciences
Stellenbosch University

Contents

1	Metric Spaces, Normed Spaces and Inner Product Spaces	1
1.1	The Geometry of \mathbb{R}^n	1
1.2	Convergence and Continuity in Metric Spaces	3
1.3	Normed Spaces	6
1.4	Inner Product Spaces	7
1.5	Linear Operators	12
1.6	Projections in Hilbert Spaces	13
2	Basic Notions of Topology	17
2.1	Countable and Uncountable Sets	17
2.2	Open Sets and the Interior Operation	21
2.3	Closed Sets and the Closure Operation	24
2.4	Compact Spaces and Sets	27
2.5	Compactness in \mathbb{R}^n	31
2.6	Convergence and Continuity	32
2.6.1	Pull-backs and Push-forwards	32
2.6.2	Topological Characterizations of Convergence and Continuity	34
2.6.3	Continuity and Compactness	35
2.7	Separable Spaces	37
2.8	The Banach Space $\mathcal{C}[a, b]$	37
2.8.1	Uniform Convergence	37
2.8.2	Compactness: Arzelà–Ascoli Theorem*	41
2.8.3	Separability: Stone–Weierstrass Theorem*	42
3	Motivation for Measure Theory	47
3.1	What is “Area”?	47
3.2	Shortcomings of the Riemann Integral	48
3.3	Motivation from Probability Theory	51
3.4	Structure of Events	52
4	Measure Spaces	55
4.1	Events and σ -algebras	55
4.2	Measures	57
4.3	Continuity Properties of Measures	60
4.3.1	Limit Operations on Sets	60
4.3.2	Limits of Sets and Measures	61

4.4	Lebesgue Measure from Coin Tossing	62
4.5	Some Probability Theory	66
4.6	Extension of Measures*	70
4.6.1	Other Families of Sets*	70
4.6.2	The Extension Theorem*	71
4.6.3	Completion of Measure Spaces*	75
4.7	Lebesgue Measure	76
4.7.1	Lebesgue measure on \mathbb{R}	76
4.7.2	Lebesgue Measure on \mathbb{R}^d	79
5	Measurable Functions and Random Variables	81
5.1	Definition of Measurable Function	81
5.2	Combinations of Measurable Functions	84
5.3	Measures and σ -algebras from Measurable Functions	88
5.4	Information	90
6	Integration	93
6.1	Definition and Basic Properties	93
6.2	Lebesgue's Dominated Convergence Theorem	98
6.3	Measure Zero	100
6.4	Chain Rule, Change of Variables	103
6.5	Riemann Integral vs. Lebesgue Integral	105
7	Differentiation	107
7.1	Bounded Linear Operators	107
7.2	The Derivative	108
7.2.1	Definition of the Derivative	108
7.2.2	The Chain Rule	112
7.2.3	Components	114
7.2.4	Partial Derivatives and the Jacobian Matrix	116
7.2.5	A Sufficient Condition for the Existence of $Df(\mathbf{x})$	118
7.2.6	The Chain Rule: Reprise	120
7.2.7	Further Manipulation Rules	121
7.2.8	Directional Derivatives	122
7.3	Taylor Theorems	124
7.3.1	Taylor's Theorem for One Variable	124
7.3.2	A Higher-Order Taylor Theorem	127
7.4	Maxima and Minima	133
7.4.1	Topological Facts about Extrema	133
7.4.2	Maxima and Minima via Calculus	134
7.4.3	Linear Least Squares	139
8	Products and Independence	141
8.1	The Monotone Class Theorem	141
8.2	Products	144
8.2.1	Introduction	144
8.2.2	Products of Measure Spaces	145

8.3	Independence	148
9	The \mathcal{L}^p-Spaces and Fourier Analysis	153
9.1	\mathcal{L}^p Spaces	153
9.1.1	Integration of complex-valued functions	153
9.1.2	Definition of \mathcal{L}^p -spaces	154
9.1.3	\mathcal{L}^1 and \mathcal{L}^2	155
9.1.4	General Theory of \mathcal{L}^p -spaces*	160
9.2	Geometry of Hilbert Space and Generalized Fourier Series	164
9.2.1	Projections in Hilbert Spaces	164
9.2.2	Orthonormal Bases	167
9.3	Fourier Series	172
9.3.1	Statement of Results	172
9.3.2	Examples	174
9.3.3	Proofs*	175
10	Weak Convergence and Characteristic Functions	181
10.1	Weak Convergence and Convergence in Distribution	181
10.2	Characteristic Functions	187
10.2.1	Basic Properties	187
10.2.2	Inversion	191
10.2.3	Weak Convergence and Characteristic Functions	196
10.3	The Central Limit Theorem	198
11	Conditional Expectation and Martingales	201
11.1	Information and Expectation	201
11.1.1	Conditioning on an Event	201
11.1.2	Conditioning on a Random Variable	203
11.1.3	Conditioning on a σ -Algebra	206
11.2	Theory of Martingales in Discrete Time	212
11.2.1	Stochastic Processes and Filtrations	212
11.2.2	Martingales, Submartingales, Supermartingales	213
11.2.3	Games and Strategies	216
12	PDEs in Finance, with a Detour Through Black-Scholes	221
12.1	Modelling Stock Prices	221
12.1.1	Modelling Returns in Continuous-Time	222
12.1.2	Modelling Share Prices in Continuous Time	225
12.2	A Naive Approach to Stochastic Calculus	226
12.3	The Black-Scholes Model	229
12.3.1	The Black-Scholes PDE	229
12.3.2	Pricing in the Risk-Neutral World	230
12.3.3	The Distribution of Asset Prices	232
12.4	Option Pricing: The Black-Scholes Formula	236

13 Introduction to PDEs	241
13.1 What is a PDE?	241
13.1.1 Types of PDEs	242
13.2 Solutions to a PDE	242
13.2.1 Contrast with ODEs	242
13.2.2 First-Order Linear PDEs	243
13.2.3 Initial- and Boundary Conditions	247
13.2.4 Well-Posed Problems	250
13.3 Classification of Linear Second-Order PDEs	251
13.4 Characteristics for 2nd-Order Linear PDEs	256
14 Laplace's Equation	263
14.1 The Divergence Theorem and Related Results	263
14.2 Harmonic Functions	265
14.2.1 Some Heuristic Remarks about the Laplace Operator	266
14.2.2 Volumes and Surface Areas of Balls in \mathbb{R}^n	267
14.2.3 Mean-Value Property and the Maximum Principle	268
14.3 Solving Laplace's Equation	272
14.3.1 Uniqueness of Solutions	272
14.3.2 Fundamental Solution of Laplace's Equation	274
14.3.3 Green's Functions	277
14.3.4 Properties of Green's functions	279
14.4 Green's Functions: Examples and Exercises	282
14.4.1 Green's function for the half-space	282
14.4.2 Green's function for the ball	285
15 The Heat Equation	289
15.1 Separation of Variables	289
15.2 The Fundamental Solution	294
15.3 Solving the Heat Equation	298
15.3.1 The Cauchy Problem	298
15.3.2 Diffusion on the Half-Line: The Method of Images	300
15.4 Applications to Finance	302
15.4.1 The Black-Scholes Option Formula	302
15.4.2 Barrier Options	303
15.5 Distributions	306
15.5.1 Basic Definitions	306
15.5.2 Convergence of Distributions	309
15.5.3 Differentiation of Distributions	310
15.6 Green's Functions Revisited	312
15.6.1 Laplace's Equation	312
15.6.2 The Heat Equation	313
16 The Radon-Nikodým Theorem*	315
16.1 Definitions and Statement of Radon-Nikodým Theorem	315
16.2 Proof of the Radon-Nikodým Theorem; Related Results	316
16.3 Products	321

16.3.1	Introduction	321
16.3.2	Products of Measure Spaces	322
A	Convergence in \mathbb{R}	327
A.1	Definition of Convergence	327
A.2	The Completeness Axiom	330
A.3	\limsup and \liminf ; Subsequences	333
A.4	Cauchy Sequences and Completeness	340
B	Sets and Logic	345
B.1	Logic, Formal Languages, Quantifiers	345
B.2	Basic Set Theory	347
B.2.1	Sets	347
B.2.2	Union and intersection	348
B.2.3	Set difference, complementation and symmetric difference	349
B.2.4	Set algebra	349
B.2.5	Products	351
B.3	The Extended Real Number System	352

Chapter 1

Metric Spaces, Normed Spaces and Inner Product Spaces

1.1 The Geometry of \mathbb{R}^n

I will assume that you are thoroughly familiar with the following facts and notions:

- \mathbb{R}^n is the set of all ordered n -tuples with components in \mathbb{R} :

$$\mathbb{R}^n = \{(r_1, \dots, r_n) : r_i \in \mathbb{R}, 1 \leq i \leq n\}$$

\mathbb{R}^1 is identified with \mathbb{R} , and called the *real line*; \mathbb{R}^2 is called the *real plane*.

\mathbb{R}^n is commonly referred to as n -dimensional *Euclidean space*, and also *Cartesian space*.

The elements of \mathbb{R}^n may also be referred to as real n -dimensional *vectors*.

- \mathbb{R}^n can be endowed with operations of *addition* and *scalar multiplication*:

$$\begin{aligned}(x_1, \dots, x_n) + (y_1, \dots, y_n) &= (x_1 + y_1, \dots, x_n + y_n) \\ \alpha(x_1, \dots, x_n) &= (\alpha x_1, \dots, \alpha x_n) \quad \alpha \in \mathbb{R}\end{aligned}$$

This makes \mathbb{R}^n into an n -dimensional vector space (over the scalar field \mathbb{R}). The vector

$$0 = (0, \dots, 0)$$

is an identity element for the operation of addition. We denote it simply by 0.

Thus we use the same symbol 0 for the number 0, the vector (0, 0), the vector (0, 0, 0), etc. Which zero is meant will be obvious from context.

- \mathbb{R}^n can be equipped with an *inner product*, a map

$$\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$$

defined as follows: If $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$, then

$$\langle x, y \rangle = x_1 y_1 + \dots + x_n y_n$$

The inner product on \mathbb{R}^1 is just ordinary multiplication. On \mathbb{R}^n , the inner product is also called the *dot product* and often denoted by $\langle x, y \rangle = x \cdot y$. It has the following properties, which you are invited to verify yourself:

- (i) $\langle x, x \rangle \geq 0$ for all $x \in \mathbb{R}$;
- (ii) $\langle x, x \rangle = 0$ if and only if $x = 0$;
- (iii) $\langle x, y \rangle = \langle y, x \rangle$ for all $x, y \in \mathbb{R}^n$;
- (iv) $\langle x, y + z \rangle = \langle x, y \rangle + \langle x, z \rangle$ for all $x, y, z \in \mathbb{R}^n$;
- (v) $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$ for all $x, y \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$.

A vector space V equipped with a map $\langle \cdot, \cdot \rangle$ which satisfies (i)–(v) is called an *inner product space*.

- The space \mathbb{R}^n can be equipped with a *norm* or length

$$\|\cdot\| : \mathbb{R}^n \rightarrow \mathbb{R}^+$$

defined by

$$\|x\| = \sqrt{\langle x, x \rangle} = \sqrt{x_1^2 + \cdots + x_n^2}$$

The norm in \mathbb{R}^1 is just the usual absolute value. The norm satisfies the following conditions:

- (i) $\|x\| \geq 0$ for all $x \in \mathbb{R}^n$;
- (ii) $\|x\| = 0$ if and only if $x = 0$;
- (iii) $\|\alpha x\| = |\alpha| \|x\|$ for all $x \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$;
- (iv) $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in \mathbb{R}^n$ (Triangle Inequality);

A vector space V equipped with a map $\|\cdot\|$ which satisfies (i)–(iv) is called a *normed vector space*. An inner product will always *induce* a norm by putting $\|x\| = \langle x, x \rangle$. However, a norm need not be induced by an inner product.

- \mathbb{R}^n can be equipped with a *metric*, or distance

$$d(\cdot, \cdot) : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^+$$

defined by

$$d(x, y) = \|x - y\| = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}$$

$d(x, y)$ is simply the distance between x and y . The metric d satisfies the following conditions:

- (i) $d(x, y) \geq 0$ for all $x, y \in \mathbb{R}^n$;
- (ii) $d(x, y) = 0$ if and only if $x = y$;
- (iii) $d(x, y) = d(y, x)$ for all $x, y \in \mathbb{R}^n$;
- (iv) $d(x, z) \leq d(x, y) + d(y, z)$ for all $x, y, z \in \mathbb{R}^n$ (Triangle Inequality);

The Triangle Inequality for the metric follows directly from the Triangle Inequality for the norm:

$$d(x, z) = \|x - z\| = \|(x - y) + (y - z)\| \leq \|x - y\| + \|y - z\| = d(x, y) + d(y, z)$$

Note that properties (i)–(iv) for d , unlike the properties for the inner product and norm, do not mention addition or scalar multiplication at all. A *set* X (not necessarily a vector space) equipped with a map d satisfying (i)–(iv) is called a *metric space*. A norm will always induce a metric by putting $d(x, y) = \|x - y\|$.

1.2 Convergence and Continuity in Metric Spaces

Many of the structures on the space \mathbb{R}^n that we discussed in the previous section have analogues in different kinds of spaces. The most basic structure that we shall need in this course is the notion of a *metric space*:

Definition 1.2.1 A *metric space* is a pair (X, d) consisting of a set X together with a map $d : X \times X \rightarrow \mathbb{R}$, called a *metric*, which satisfies the following conditions:

- (i) $d(x, y) \geq 0$ for all $x, y \in X$;
- (ii) $d(x, y) = 0$ if and only if $x = y$;
- (iii) $d(x, y) = d(y, x)$ for all $x, y \in X$;
- (iv) $d(x, z) \leq d(x, y) + d(y, z)$ for all $x, y, z \in X$ (Triangle Inequality);

If (X, d) is a metric space, and x, y are *points* in X , then $d(x, y)$ should be interpreted as the *distance* between x and y . With this interpretation in mind, parts (i)-(iv) of Defn. 1.2.1 can be easily understood. Though we have distilled the definition of *metric* from our experience with distances in \mathbb{R}^n , there are many metric spaces which do not even remotely resemble \mathbb{R}^n . For example:

Example 1.2.2 (1) Let X be *any*¹ non-empty set, and define

$$d : X \times X \rightarrow \mathbb{R} : (x, y) \mapsto \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$$

Then (X, d) is a metric space.

Here d is called the *discrete metric*, and (X, d) is called a *discrete space*.

- (2) Let (X, d) be a metric space, and let $Y \subseteq X$. Let d_Y be the restriction of d to Y , i.e. d_Y is defined only on $Y \times Y$, and coincides there with d . Then (Y, d_Y) is a metric space. Now even though (X, d) might be “nice”, (Y, d_Y) could be quite “wild”. For example, we could take (X, d) to be the set of reals with the usual metric $d(x_0, x_1) := |x_0 - x_1|$, and let Y be the set of irrational numbers.

□

Thus many familiar operations on \mathbb{R}^n may not be possible on an abstract metric space. Addition and multiplication may be undefined, there may be no order relation, etc. But one operation is possible: It is possible to define *limits*. For once we have a *distance*, we can talk about convergence: x_n converges to x iff the distance $d(x_n, x)$ between x_n and x converges to 0. Thus convergence of a sequence of abstract points $\langle x_n \rangle_n$ is defined in terms of convergence of a sequence of real numbers $\langle d(x_n, x) \rangle_n$ — and we already know what we mean by convergence of real numbers!

Definition 1.2.3 Let (X, d) be a metric space, and let $x_n, x \in X$ (for $n \in N$). We say that

$$x_n \rightarrow x \quad \text{iff} \quad d(x_n, x) \rightarrow 0$$

Thus

$$x_n \rightarrow x \quad \text{iff} \quad \forall \varepsilon > 0 \exists N \forall n \geq N [d(x_n, x) < \varepsilon]$$

We write $\lim_n x_n = x$ when $x_n \rightarrow x$.

¹ X might be a set of cows, for example.

- Exercise 1.2.4** (a) Show that a sequence in a metric space can have at most one limit.
 (b) Show that if a sequence $\langle x_n \rangle_n$ converges in a metric space, then every subsequence of $\langle x_n \rangle_n$ converges as well, and to the same limit.

□

Once we have distance, we can also define the notion of *continuity*. Intuitively, a function f is continuous at a point x_0 iff whenever x is “very close” to x_0 , then $f(x)$ is “very close” to $f(x_0)$. Slightly more precisely,

We can make $f(x)$ as “close” to $f(x_0)$ as we wish, by taking x sufficiently “close” to x_0 .

There are no “jumps” then, because the closer x gets to x_0 , the closer $f(x)$ gets to $f(x_0)$. This is not very precise, so we need to proceed with care:

- Continuity makes sense for a function $f : X \rightarrow Y$ between two metric spaces (X, d_X) and (Y, d_Y) . We can use d_X to talk about “closeness” of x, x_0 in X , and d_Y to talk about “closeness” of $f(x), f(x_0)$ in Y .
- Since “closeness” is subjective, we will demand that it holds for absolutely anybody’s idea of “close”. Specifically, suppose you define “close” by specifying some real number $\varepsilon > 0$ and saying “ $f(x), f(x_0)$ are “close” in Y whenever $d_Y(f(x), f(x_0)) < \varepsilon$ ”. Intuitively, then, f is continuous at x_0 iff whenever x is “sufficiently close” to x_0 , then $d_Y(f(x), f(x_0)) < \varepsilon$.
- This “sufficiently close” can now be described by some real number $\delta > 0$: f is continuous at x_0 if there is $\delta > 0$ such that

$$d_Y(f(x), f(x_0)) < \varepsilon \quad \text{whenever} \quad d_X(x, x_0) < \delta$$

- Since this must hold for anyone’s idea of “closeness”, i.e. for any $\varepsilon > 0$, we now make the following definition:

Definition 1.2.5 Suppose that $f : (X, d_X) \rightarrow (Y, d_Y)$ is a function between metric spaces, and let $x_0 \in X$. We say that f is *continuous at x_0* iff

$$\forall \varepsilon > 0 \exists \delta > 0 \forall x \in X \left[d_X(x, x_0) < \delta \longrightarrow d_Y(f(x), f(x_0)) < \varepsilon \right]$$

We say that $f : (X, d_X) \rightarrow (Y, d_Y)$ is *continuous* iff it is continuous at every $x_0 \in X$.

Note that the notion of continuity depends on both the metric of the domain space and the metric of the range space.

Here is a characterization of continuity in terms of convergence of sequences.

Proposition 1.2.6 A map $f : (X, d_X) \rightarrow (Y, d_Y)$ is continuous at $x_0 \in X$ iff

$$\text{Whenever } x_n \rightarrow x_0 \text{ in } X, \text{ then } f(x_n) \rightarrow f(x_0) \text{ in } Y$$

Exercise 1.2.7 We prove Propn. 1.2.6. Let $f : (X, d_X) \rightarrow (Y, d_Y)$, and let $x_0 \in X$.

- (a) Suppose that f is continuous at x_0 , and assume $x_n \rightarrow x_0$ in (X, d_X) . We must show that $f(x_n) \rightarrow f(x_0)$ in (Y, d_Y) . So let $\varepsilon > 0$. Explain why there is $\delta > 0$ so that $d_Y(f(x), f(x_0)) < \varepsilon$ whenever $d_X(x, x_0) < \delta$. Also explain why there is $N \in \mathbb{N}$ so that $d_X(x_n, x_0) < \delta$ whenever $n \geq N$. Conclude that

$$d_Y(f(x_n), f(x_0)) < \varepsilon \quad \text{whenever} \quad n \geq N$$

and explain why this implies that $f(x_n) \rightarrow f(x_0)$ in (Y, d_Y) .

- (b) To prove the converse, suppose that f is *not* continuous at x_0 . By examining the definition of continuity, explain why there is a $\varepsilon > 0$ with the following property:

$$\forall \delta > 0 \exists x \in X \left[d_X(x, x_0) < \delta \wedge d_Y(f(x), f(x_0)) \geq \varepsilon \right]$$

Now take $\delta := \frac{1}{n}$ to find $x_n \in X$ with the property that $d_X(x_n, x_0) < \frac{1}{n}$, yet $d_Y(f(x_n), f(x_0)) \geq \varepsilon$. Conclude that $x_n \rightarrow x_0$ in X , yet $f(x_n) \not\rightarrow f(x_0)$ in Y . Explain why this proves the converse.

□

Because an abstract metric space does not come with an order relation, we cannot define sup and inf in an arbitrary metric space. Hence the Completeness Axiom makes no sense in a metric space. We can, however, use the equivalence suggested by Exercise A.4.7:

Definition 1.2.8 Let (X, d) be a metric space.

- (a) A sequence $\langle x_n \rangle_n$ in X is called a *Cauchy sequence* if and only if

$$\sup_{m, n \geq N} d(x_m, x_n) \rightarrow 0 \quad \text{as } N \rightarrow \infty$$

i.e. iff for every $\varepsilon > 0$ there is an $N \in \mathbb{N}$ such that

$$d(x_n, x_m) < \varepsilon \quad \text{whenever} \quad n, m > N$$

- (b) We say that (X, d) is *complete* iff every Cauchy sequence in X converges (to a limit which is in X).

Exercise 1.2.9 (a) Show that a convergent sequence is always a Cauchy sequence, in any metric space.

(b) Show that a discrete metric space (cf. Example 1.2.2(1)) is necessarily complete.

(c) Show that there exist metric spaces which are not complete, i.e. in which not all Cauchy sequences converge.

□

It is possible to define convergence and continuity in spaces which are even more abstract than metric spaces, using even vaguer notions of “closeness” which do not depend on having a distance function (i.e. a metric). In these so-called *topological spaces*, the fundamental notion is that of an *open set*. You will shortly get see how this works. In many cases, however, the spaces in which we work have *more* structure, rather than less, and it is to these that we now turn.

1.3 Normed Spaces

Throughout this section we consider vector spaces V over the scalar field \mathbb{R} .²

Definition 1.3.1 A *normed space* is a pair $(V, \|\cdot\|)$, where V be a vector space and $\|\cdot\|$ is a *norm* on V , i.e. a function $\|\cdot\| : V \rightarrow \mathbb{R}$ with the following properties:

- (i) $\|x\| \geq 0$ for all $x \in V$;
- (ii) $\|x\| = 0$ if and only if $x = 0$;
- (iii) $\|\alpha x\| = |\alpha| \|x\|$ for all $x \in V$ and $\alpha \in \mathbb{R}$;
- (iv) $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in V$ (Triangle Inequality);

The norm $\|v\|$ of a vector $v \in V$ should be interpreted as its *length*, as in the following standard examples.

Examples 1.3.2 (a) $V := \mathbb{R}$ with $\|v\| := |v|$ (the absolute value).

(b) $V := \mathbb{R}^n$ with the *Euclidean norm*

$$\|\mathbf{v}\| := \left(\sum_{k=1}^n v_k^2 \right)^{\frac{1}{2}} \quad \text{where } \mathbf{v} = (v_1, \dots, v_n)$$

□

Here are some other norms on \mathbb{R}^n :

Exercise 1.3.3 For $\mathbf{x} \in \mathbb{R}^n$, let $\mathbf{x} = (x_1, \dots, x_n)$

(a) Define $\|\cdot\|_1$ on \mathbb{R}^n by

$$\|\mathbf{x}\|_1 = |x_1| + \dots + |x_n|$$

Show that $\|\cdot\|_1$ is a norm on \mathbb{R}^1 .

(b) Define $\|\cdot\|_\infty$ on \mathbb{R}^n by

$$\|\mathbf{x}\|_\infty = \max\{|x_1|, \dots, |x_n|\}$$

Show that $\|\cdot\|_\infty$ is a norm on \mathbb{R}^n ;

□

We will give more interesting examples shortly.

Exercise 1.3.4 Let $(V, \|\cdot\|)$ be a normed space.

(a) Prove that $\|x\| = \|-x\|$ for all $x \in V$.

(b) Prove that

$$\|x - y\| \geq \left| \|x\| - \|y\| \right| \quad \text{for all } x, y \in V$$

(c) Prove that the norm-mapping is continuous, i.e that the map $V \rightarrow \mathbb{R} : x \mapsto \|x\|$ is continuous.

□

Conditions (i)-(iv) in the definition of a norm seem similar to the conditions (i)-(iv) in the definition of a metric. This is no accident:

²Later in this course, we may have occasion to consider vector spaces over \mathbb{C} as well.

Proposition 1.3.5 *Every norm induces a metric: If $(V, \|\cdot\|)$ is a normed space, then (V, d) is a metric space, where $d : V \times V \rightarrow \mathbb{R}$ is defined by*

$$d(v, w) := \|v - w\|$$

Exercise 1.3.6 (a) Prove Propn. 1.3.5.

(b) Prove that the converse of Propn. 1.3.5 is false: Not every metric on a vector space V is obtained from a norm.

[Hint: Consider the discrete metric on a vector space V .]

□

Here is a first look at *function spaces*:

Exercise 1.3.7 Suppose that $[a, b]$ is a closed interval in \mathbb{R} . Let $\mathcal{C}[a, b]$ be the set of all continuous functions $f : [a, b] \rightarrow \mathbb{R}$.

(a) Show that $\mathcal{C}[a, b]$ is a vector space, when the operations of addition and scalar multiplication are defined pointwise.

(b) Define $\|\cdot\|_1 : \mathcal{C}[a, b] \rightarrow \mathbb{R}$ by

$$\|f\|_1 := \int_a^b |f(t)| dt$$

Show that $\|\cdot\|_1$ is a norm on $\mathcal{C}[a, b]$.

(c) Define $\|\cdot\|_\infty : \mathcal{C}[a, b] \rightarrow \mathbb{R}$ by

$$\|f\|_\infty := \sup\{|f(t)| : t \in [a, b]\}$$

Show that $\|\cdot\|_\infty$ is a norm on $\mathcal{C}[a, b]$.

□

Exercise 1.3.8 Let l^1 be the set of all sequences $\langle x_n \rangle_n$ in \mathbb{R} with the property that $\sum_n |x_n| < \infty$. Show that l^1 is a vector space, and that $\|\langle x_n \rangle_n\|_1 := \sum_n |x_n|$ defines a norm on l^1 .

□

1.4 Inner Product Spaces

Next, we consider an additional structure on a real vector space V :

Definition 1.4.1 An *inner product space* is a pair $(V, \langle \cdot, \cdot \rangle)$, where V be a vector space and $\langle \cdot, \cdot \rangle$ is an *inner product* on V , i.e. a function $\langle \cdot, \cdot \rangle : V \times V \rightarrow \mathbb{R}$ with the following properties:

- (i) $\langle x, x \rangle \geq 0$ for all $x \in V$;
- (ii) $\langle x, x \rangle = 0$ if and only if $x = 0$;
- (iii) $\langle x, y \rangle = \langle y, x \rangle$ for all $x, y \in V$;
- (iv) $\langle x, y + z \rangle = \langle x, y \rangle + \langle x, z \rangle$ for all $x, y, z \in V$;
- (v) $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$ for all $x, y \in V$ and $\alpha \in \mathbb{R}$.

Example 1.4.2 \mathbb{R}^n is an inner product space when equipped with the usual *dot product*:

$$\langle \mathbf{x}, \mathbf{y} \rangle := \mathbf{x} \cdot \mathbf{y} = \sum_{j=1}^n x_j y_j$$

where $\mathbf{x} = (x_1, \dots, x_n), \mathbf{y} = (y_1, \dots, y_n)$.

□

We shall shortly present more interesting examples of inner product spaces. For the moment, we note that every inner product induces a norm, in the same way that the dot product in \mathbb{R}^n yields a length: The length of a vector $\mathbf{x} \in \mathbb{R}^n$ is given by $\sqrt{\mathbf{x} \cdot \mathbf{x}}$, and it turns out that $\|x\| := \sqrt{\langle x, x \rangle}$ defines a norm in terms of the inner product.

To prove that, we need the following result:

Proposition 1.4.3 (Cauchy–Schwarz Inequality)^a

If $(V, \langle \cdot, \cdot \rangle)$ is an inner product space, then

$$|\langle x, y \rangle| \leq \sqrt{\langle x, x \rangle} \sqrt{\langle y, y \rangle}$$

for all $x, y \in V$.

Moreover we have equality iff x is a scalar multiple of y .

^aAlso called the Cauchy–Bunyakovskii–Schwarz Inequality

Exercise 1.4.4 We prove Propn. 1.4.3. Let $(V, \langle \cdot, \cdot \rangle)$ be an inner product space.

(a) First show that for $\alpha \in \mathbb{R}$ and $x, y \in V$ we have

$$0 \leq \langle x - \alpha y, x - \alpha y \rangle = \langle x, x \rangle - 2\alpha \langle x, y \rangle + \alpha^2 \langle y, y \rangle$$

(b) Now, with x, y held fixed, consider the righthand side of the above inequality as a *quadratic polynomial* in α . By examining its discriminant, explain why

$$\langle x, y \rangle^2 - \langle x, x \rangle \langle y, y \rangle \leq 0$$

with equality only when $x = \alpha y$ for some α .

(c) Now conclude the result.

□

Proposition 1.4.5 If $(V, \langle \cdot, \cdot \rangle)$ is an inner product space, then the map $\|\cdot\| : V \rightarrow \mathbb{R}$ given by

$$\|x\| := \sqrt{\langle x, x \rangle}$$

defines a norm on V .

Exercise 1.4.6 Prove Propn. 1.4.5.

□

Exercise 1.4.7 (a) Suppose that $(V, \langle \cdot, \cdot \rangle)$ is an inner product space, and that $\|\cdot\|$ is the norm induced by the inner product. Prove that if $x, y \in V$, then

$$\|x + y\|^2 + \|x - y\|^2 = 2(\|x\|^2 + \|y\|^2)$$

Why do you think this identity is called the *Parallelogram Law*?

- (b) **Prove or Disprove:** If $(V, \|\cdot\|)$ is a normed space, then it is possible to define an inner product $\langle \cdot, \cdot \rangle : V \times V \rightarrow \mathbb{R}$ such that $\|\cdot\|$ is the norm induced by this inner product.
- (c) **Prove or Disprove:** If $(V, \|\cdot\|)$ is a normed space which satisfies the Parallelogram Law, then it is possible to define an inner product $\langle \cdot, \cdot \rangle : V \times V \rightarrow \mathbb{R}$ such that $\|\cdot\|$ is the norm induced by this inner product.

□

In \mathbb{R}^n , the dot product does not only induce a length; it also induces an *angle*: The angle θ between two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ is given by

$$\cos \theta = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

We can imitate this definition in an abstract inner product space $(V, \langle \cdot, \cdot \rangle)$, and define the angle between $x, y \in V$ by

$$\cos \theta := \frac{\langle x, y \rangle}{\|x\| \|y\|} \quad \text{where } \|x\| := \sqrt{\langle x, x \rangle}$$

By the Cauchy–Schwarz inequality it follows immediately that $|\cos \theta| \leq 1$, so that this definition makes sense. It also follows that $|\cos \theta| = 1$ if and only if x is a scalar multiple of y , i.e. iff x, y are parallel. We can also define *orthogonality* in an abstract inner product space, in the obvious way:

Definition 1.4.8 Suppose that $(V, \langle \cdot, \cdot \rangle)$ is an inner product space. We say that $x, y \in V$ are *orthogonal*, and write $x \perp y$, if and only if $\langle x, y \rangle = 0$.
If $G \subseteq V$, we say that $x \perp G$ iff $\forall g \in G (x \perp g)$.

The following exercise gives a nice example of an inner product space that we shall meet again later. It shows that several commonly used statistical terms are actually derived from inner product spaces.

Exercise 1.4.9 Consider a random experiment — because we do not yet have the measure theoretic machinery required, we have to be a little imprecise here — and let V be the set of all random variables with zero mean and finite variance.

- (a) Show that V is a vector space.
[Hint: Imitate the proof of the Cauchy–Schwarz inequality to show that if $X, Y \in V$, then $X + Y$ has finite variance.]
- (b) Define $\langle \cdot, \cdot \rangle : V \times V \rightarrow \mathbb{R}$ by $\langle X, Y \rangle = \mathbb{E}[XY]$ (where $\mathbb{E}X$ denotes the expectation of X). Show that $\langle \cdot, \cdot \rangle$ is an inner product on V .
- (c) This inner product induces a norm $\|\cdot\|$. What is the usual name for $\|X\|$ in statistics?
- (d) The inner product also induces an angle θ between two random variables $X, Y \in V$. What is the usual name for $\cos \theta$ in statistics?

□

Exercise 1.4.10 Let $\mathcal{C}[a, b]$ be the set of all continuous functions $f : [a, b] \rightarrow \mathbb{R}$. We already know from Exercise 1.3.7 that $\mathcal{C}[a, b]$ is a vector space, and we defined two norms $\|\cdot\|_1$ and $\|\cdot\|_\infty$ on this space.

- (a) Show that the Parallelogram Law fails for both $\|\cdot\|_1$ and $\|\cdot\|_\infty$. Thus these norms are not induced by an inner product.

(b) Define a map $\langle \cdot, \cdot \rangle : \mathcal{C}[a, b] \times \mathcal{C}[a, b] \rightarrow \mathbb{R}$ by

$$\langle f, g \rangle := \int_a^b f(t)g(t) dt$$

Show that $\langle \cdot, \cdot \rangle$ defines an inner product on $[a, b]$.

(c) The induced norm is therefore

$$\|f\|_2 := \left(\int_a^b f(t)^2 dt \right)^{\frac{1}{2}}$$

Compare the norms $\|\cdot\|_1$, $\|\cdot\|_2$ and $\|\cdot\|_\infty$ with the norms on \mathbb{R}^n discussed in Example 1.3.2 and Exercise 1.3.3.

□

Once we have introduced some measure theory, it will become apparent that Exercises 1.4.9 and 1.4.10 are really instances of the same idea.

In \mathbb{R}^n , the *standard basis* $\{\mathbf{e}_i : i = 1, \dots, n\}$ is *orthonormal* i.e.

$$\langle \mathbf{e}_i, \mathbf{e}_j \rangle = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases} \quad \begin{matrix} \text{(orthogonal)} \\ \text{(normal)} \end{matrix}$$

Definition 1.4.11 Let $(V, \langle \cdot, \cdot \rangle)$ be an inner product space and let $U \subseteq V$. We say that U is *orthonormal* iff for all $u, u' \in U$, we have

$$\langle u, u' \rangle = \begin{cases} 0 & \text{if } u \neq u' \\ 1 & \text{if } u = u' \end{cases}$$

Exercise 1.4.12 If U is an orthonormal set, then the vectors in U are mutually linearly independent. [Hint: Suppose that $\sum_k c_k u_k = 0$ and consider $\langle u_j, 0 \rangle$.]

□

Now assume that $(V, \langle \cdot, \cdot \rangle)$ is an inner product space which possesses an *orthonormal basis* $\{u_1, \dots, u_n\}$ — so that V is finite-dimensional. It is then very easy to represent every vector in V as a linear combination of these basis vectors: If $v = \sum_{k=1}^n c_k u_k$, then $\langle v, u_j \rangle = \sum_{k=1}^n c_k \langle u_k, u_j \rangle = c_j$, and so

$$v = \sum_{k=1}^n \langle v, u_k \rangle u_k$$

The next theorem shows that this can always be done in a finite dimensional inner product space:

Theorem 1.4.13 (Gram–Schmidt Orthogonalization) *If $(V, \langle \cdot, \cdot \rangle)$ is a finite dimensional inner product space, then V has an orthonormal basis.*

Proof: Suppose that $\{v_1, \dots, v_n\}$ is a basis of V . We proceed by inductively building an orthonormal basis $\{u_1, \dots, u_n\}$ so that

$$\text{span}\{v_1, \dots, v_i\} = \text{span}\{u_1, \dots, u_i\} \quad \text{for } i = 1, \dots, n$$

Define $u_1 := \frac{v_1}{\|v_1\|}$. Then $\|u_1\| = 1$, en $\text{span}\{u_1\} = \text{span}\{v_1\}$.

Assume now that we have already defined u_1, \dots, u_i (for $1 \leq i < n$), so that $\{u_1, \dots, u_i\}$ is an orthonormal set with the same span as $\{v_1, \dots, v_i\}$. We must now define u_{i+1} . First define

$$w_{i+1} = v_{i+1} - \sum_{j=1}^i \langle v_{i+1}, u_j \rangle u_j$$

and note that

(i) $w_{i+1} \neq 0$, for otherwise 0 would be a linear combination of u_1, \dots, u_i and v_i , and thus a linear combination of v_1, \dots, v_{i+1} . But v_1, \dots, v_{i+1} is linearly independent.

(ii) If $1 \leq j \leq i$, then

$$\langle w_{i+1}, u_j \rangle = \langle v_{i+1}, u_j \rangle - \sum_{k=1}^n \langle v_{i+1}, u_k \rangle \langle u_k, u_j \rangle = 0$$

It follows that u_1, \dots, u_i, w_{i+1} is an orthogonal set, and thus linearly independent.

(iii) As $\text{span}\{u_1, \dots, u_i\} = \text{span}\{v_1, \dots, v_i\}$, we see that $\text{span}\{u_1, \dots, u_i, w_{i+1}\} = \text{span}\{v_1, \dots, v_{i+1}\}$.

The only potential problem is that we might not have $\|w_{i+1}\| = 1$. Therefore, define

$$u_{i+1} := \frac{w_{i+1}}{\|w_{i+1}\|}$$

.

+

The next exercise shows that in a finite-dimensional inner product space, inner products are essentially just dot products:

Exercise 1.4.14 Suppose that $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ is an orthonormal basis for V . If $\mathbf{v} = \sum_{k=1}^n v_k \mathbf{u}_k$ and $\mathbf{w} = \sum_{k=1}^n w_k \mathbf{u}_k$, then

- (a) $\langle \mathbf{v}, \mathbf{w} \rangle = \sum_{k=1}^n v_k w_k$
- (b) $\langle \mathbf{v}, \mathbf{w} \rangle = \sum_{k=1}^n \langle \mathbf{v}, \mathbf{u}_k \rangle \langle \mathbf{u}_k, \mathbf{w} \rangle$
- (c) $\|\mathbf{v}\| = (\sum_{k=1}^n \langle \mathbf{v}, \mathbf{u}_k \rangle^2)^{\frac{1}{2}}$.

□

At a later stage, we will encounter these ideas — orthonormal representations of vectors in inner product spaces — again: in infinite-dimensional function spaces, when we discuss Fourier analysis.

1.5 Linear Operators

Suppose that V, W are vector spaces. A linear operator $T : V \rightarrow W$ is one which preserves the operations of addition and scalar multiplication, i.e. one which satisfies

$$T(x + y) = Tx + Ty \quad T(\alpha x) = \alpha Tx$$

We now tackle the question of when a linear operator is *continuous*. To do that, we require additional structure on V and W , because continuity involves the idea of “closeness”. We will henceforth consider linear operators between normed spaces.

By translating the definition of continuity from the language of metrics to the language of norms, we see that a map $T : (V, \|\cdot\|_V) \rightarrow (W, \|\cdot\|_W)$ is continuous if and only if

$$\forall x \in V \forall \varepsilon > 0 \exists \delta > 0 \forall y \in V [\|x - y\|_V < \delta \Rightarrow \|Tx - Ty\|_W < \varepsilon]$$

When T is linear, we have $Tx - Ty = T(x - y)$. Setting $h := x - y$, we thus obtain

$$\|h\|_V < \delta \Rightarrow \|Th\|_W < \varepsilon$$

This suggests the following definition:

Definition 1.5.1 A linear operator $T : (V, \|\cdot\|_V) \rightarrow (W, \|\cdot\|_W)$ between normed vector spaces is said to be *bounded* if and only if there is a constant $C \in \mathbb{R}$ such that $\|Tx\|_W \leq C\|x\|_V$ for all $x \in V$.

To simplify notation, we will use the same symbol $\|\cdot\|$ for $\|\cdot\|_V, \|\cdot\|_W$, etc. This should not cause any confusion; you need merely look where the vectors reside. Thus if $T : V \rightarrow W$ and $x \in V$, then $\|x\| = \|x\|_V$ and $\|Tx\| = \|Tx\|_W$.

The next lemma shows that there are plenty of bounded linear operators:

Lemma 1.5.2 Every linear operator defined on a Euclidean space is bounded, i.e. if $T : (\mathbb{R}^n, \|\cdot\|) \rightarrow (V, \|\cdot\|)$ is a linear transformation, then there exists a constant C such that

$$\|Tx\| \leq C \|x\| \quad \text{for all } x \in \mathbb{R}^n$$

(where the norm on \mathbb{R}^n is the standard Euclidean norm.)

Proof: Let $\mathbf{e}_1, \dots, \mathbf{e}_n$ denote the standard basis of \mathbb{R}^n , and put $C = n \max\{\|T\mathbf{e}_1\|, \dots, \|T\mathbf{e}_n\|\}$, so that each $\|T\mathbf{e}_i\| \leq \frac{C}{n}$. If $h = (h_1, \dots, h_n)^{tr} \in \mathbb{R}^n$, then $h = \sum_{i=1}^n h_i \mathbf{e}_i$, and each $|h_i| = \sqrt{h_i^2} \leq \sqrt{h_1^2 + \dots + h_n^2} = \|h\|$. It follows, using the triangle inequality and the inequalities just obtained that

$$\|Th\| = \|T(\sum_{i=1}^n h_i \mathbf{e}_i)\| \leq \sum_{i=1}^n |h_i| \|T\mathbf{e}_i\| \leq \sum_{i=1}^n \|h\| \frac{C}{n} = C \|h\|$$

—

The next proposition shows that a linear operator is continuous if and only if it is bounded:

Proposition 1.5.3 Let $T : V \rightarrow W$ be a linear operator between normed vector spaces V, W . Then T is continuous if and only if it is bounded,

Proof: Suppose T is continuous. Choose $\delta > 0$ so that $\|Tx\| < 1$ whenever $\|x\| < \delta$. Let $C \geq \frac{1}{\delta}$. If $x \in V$, then $\|\frac{x}{C\|x\|}\| < \delta$, so $\|T(\frac{x}{C\|x\|})\| < 1$, i.e. $\|Tx\| \leq C\|x\|$.

Conversely, suppose that T is a bounded operator, and that $\|Tx\| \leq C\|x\|$ for all $x \in V$. To show that T is continuous, it suffices to show that $Tx_n \rightarrow Tx$ whenever $x_n \rightarrow x$, i.e. that $\|Tx_n - Tx\| \rightarrow 0$ whenever $\|x_n - x\| \rightarrow 0$. But this is easy:

$$\|Tx_n - Tx\| = \|T(x_n - x)\| \leq C\|x_n - x\| \rightarrow 0 \text{ as } \|x_n - x\| \rightarrow 0$$

□

Proposition 7.1.1 and Lemma 7.1.2 immediately imply that:

Corollary 1.5.4 Any linear operator $L : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is continuous.

Remarks 1.5.5 Since every finite dimensional real vector space is isomorphic to a Euclidean space, the preceding corollary proves that all linear operators defined on finite dimensional normed vector spaces are continuous. This breaks down for infinite dimensional vector spaces, as you will see in the next exercise.

□

Exercise 1.5.6 Consider the subspace $\mathcal{C}^1[a, b] \subseteq \mathcal{C}[a, b]$ which consists of all continuously differentiable³ functions $f : [a, b] \rightarrow \mathbb{R}$. Equip $\mathcal{C}^1[a, b]$ with the $\|\cdot\|_\infty$ -norm, i.e. $\|f\|_\infty := \sup_{t \in [a, b]} |f(t)|$. For $t_0 \in (a, b)$, define the map $D_{t_0} : \mathcal{C}^1[a, b] \rightarrow \mathbb{R}$ by

$$D_{t_0}f = f'(t_0)$$

(a) Show that D_{t_0} is linear.

(b) Show that D_{t_0} is *not* generally continuous.

[Hint: Define $f_n(x) := \frac{1}{n} \sin 2\pi nx$. Then $f_n \rightarrow 0$. Take $t_0 = \frac{1}{2}$.]

□

1.6 Projections in Hilbert Spaces

We end this chapter with an important concept: that of *orthogonal projection*.

If V is a linear subspace of \mathbb{R}^n , then we can project any $\mathbf{x} \in \mathbb{R}^n$ onto V . That is, we can represent \mathbf{x} as a sum

$$\mathbf{x} = \mathbf{x}^\parallel + \mathbf{x}^\perp \quad \text{where } \mathbf{x}^\parallel \in V, \quad \mathbf{x}^\perp \perp V$$

One can think of \mathbf{x}^\parallel as the best approximation to \mathbf{x} in V : It is the vector in V which lies closest to \mathbf{x} .

This can also be done in an inner product space, provided that the subspace is *complete*. Recall that a metric space is said to be complete if and only if every Cauchy sequence converges. Since every inner product induces a norm, and since every norm induces a metric, it makes sense to talk about complete inner product spaces and complete normed spaces. Specifically: A sequence $\langle v_n \rangle$ in a normed space $(V, \|\cdot\|)$ is a Cauchy sequence iff $\sup_{n, m \geq N} \|v_n - v_m\| \rightarrow 0$ as $N \rightarrow \infty$. A normed space $(V, \|\cdot\|)$ is complete iff every Cauchy sequence in V converges (to a vector in V).

Complete spaces are important enough to warrant names:

- A complete normed space is called a *Banach space*.

³i.e the derivative exists and is continuous.

- A complete inner product space is called a *Hilbert space*.

A subspace W of a Banach (Hilbert) space $(V, \langle \cdot, \cdot \rangle)$ need not be itself a Banach (Hilbert) space. If it is, the subspace is said to be *closed*.

Example 1.6.1 We will shortly see that the space $(\mathcal{C}[-1, 1], \|\cdot\|_\infty)$ is a Banach space. The absolute value function $f(x) := |x|$ can be approximated uniformly by continuously differentiable functions, i.e. there exists a sequence f_n of continuously differentiable functions such that $\|f_n - f\|_\infty \rightarrow 0$. Thus the space $\mathcal{C}^1[-1, 1]$ of continuously differentiable functions is a subspace of $\mathcal{C}[-1, 1]$ which is *not* closed: $\langle f_n \rangle_n$ is a Cauchy sequence in $\mathcal{C}^1[-1, 1]$ which does not converge to a point (function) in $\mathcal{C}^1[-1, 1]$. □

Suppose that V is a Hilbert space, and that W is a closed linear subspace of V . If $v_0 \in V$, we can find the *best approximation* of v_0 in W . This is the unique vector w_0 with the properties that

- (i) $w_0 \in W$, and
- (ii) $\|v_0 - w_0\| = \inf\{\|v_0 - w\| : w \in W\}$, i.e. w_0 is the vector in W that lies closest to v_0 .
- (iii) Moreover, $(v_0 - w_0) \perp W$.

The vector w_0 satisfying (i)–(iii) is called the *orthogonal projection of v_0 onto W* . Indeed, $v_0 = w_0 + (v_0 - w_0)$ decomposes v_0 into a vector in W and a vector orthogonal to W . It remains to show that orthogonal projections exist and are unique.

Proposition 1.6.2 *Let V be a Hilbert space, and let W be a closed linear subspace of V . Then any v_0 in V has a unique decomposition*

$$v_0 = v_0^\parallel + v_0^\perp \quad \text{where } v_0^\parallel \in W, \quad v_0^\perp \perp W$$

v_0^\parallel is called the orthogonal projection of v_0 onto W .

Proof: *Uniqueness:* If

$$v_0 = v_0^\parallel + v_0^\perp = u_0^\parallel + u_0^\perp$$

where $v_0^\parallel, u_0^\parallel \in W$ and $v_0^\perp, u_0^\perp \perp W$, then

$$v_0^\parallel - u_0^\parallel = u_0^\perp - v_0^\perp =: x$$

is a vector with the properties that $x \in W$ and that $x \perp W$. This implies that $x \perp x$, i.e. that $\langle x, x \rangle = 0$. Hence $x = 0$, and so $v_0^\parallel = u_0^\parallel$, $v_0^\perp = u_0^\perp$.

Existence: Let $\delta = \inf\{\|v_0 - w\| : w \in W\}$, and choose a sequence $w_n \in W$ such that $\|v_0 - w_n\| \rightarrow \delta$. We show that $(w_n)_n$ is a Cauchy sequence in W : for if $\varepsilon > 0$, we may choose N such that $\|v_0 - w_n\|^2 - \delta^2 < \varepsilon$ whenever $n \geq N$. By the Parallelogram Law it follows that if $n, m \geq N$, then

$$2\varepsilon + 2\delta^2 > \|v_0 - w_n\|^2 + \|v_0 - w_m\|^2 = 2\|v_0 - \frac{1}{2}(w_n + w_m)\|^2 + 2\|\frac{1}{2}(w_n - w_m)\|^2 \geq 2\delta^2 + \frac{1}{2}\|w_n - w_m\|^2$$

Since $(w_n)_n$ is a Cauchy sequence, and since W is closed, there is $w_0 \in W$ such that $w_n \rightarrow w_0$. We will show that $w_0 = v_0^\parallel$. The fact that $\|v_0 - w_0\| \leq \|v_0 - w_n\| + \|w_n - w_0\|$ (for all $n \in \mathbb{N}$) then is easily seen to imply that $\|v_0 - w_0\| = \delta$.

It remains to show that $v_0 - w_0 \perp W$. Given an arbitrary $w \in W$ and $\lambda \in \mathbb{R}$, have $\|v_0 - w_0\|^2 = \delta^2 \leq \|v_0 - (w_0 + \lambda w)\|^2$, so that

$$-2\lambda\langle v_0 - w_0, w \rangle + \lambda^2\|w\|^2 \geq 0$$

Since this holds for all λ we must have $\langle v_0 - w_0, w \rangle = 0$. (Another way to see this is to note that the quadratic in λ has a unique root at $\lambda = 0$) and to calculate the discriminant.)

—

Remarks 1.6.3 An examination of the proof above shows that we require only that W is complete, i.e. if W is a complete subspace of an inner product space V , then orthogonal projections onto W exist.

□

Exercise 1.6.4 Suppose that W is a closed subspace of a Hilbert space $(V, \langle \cdot, \cdot \rangle)$. For each $v \in V$, let v^\parallel be the orthogonal projection of v onto W . Define a map $P : V \rightarrow V$ by $Pv := v^\parallel$. Use the uniqueness of the orthogonal projection to prove the following results.

- (a) Show that P is a bounded linear operator.
- (b) Show that P is *idempotent*, i.e. that $P^2 = P$ (i.e. $P(Pv) = Pv$ for all $v \in V$).
- (c) Show that P is *self-adjoint*, i.e. that $\langle Pv_1, v_2 \rangle = \langle Pv_1, Pv_2 \rangle = \langle v_1, Pv_2 \rangle$ for all $v_1, v_2 \in V$.
- (d) Show that $\ker P = (\text{ran } P)^\perp = W^\perp$.

□

Remarks 1.6.5 A bounded linear operator $P : V \rightarrow V$ satisfying (b), (d) in Exercise 1.6.4 is called a *projection*. The exercise shows that every orthogonal projection is a projection. It can be shown that every projection P is an orthogonal projection, onto the space $\text{ran } P$.

□

We will encounter these ideas again later, when we discuss *conditional expectation* of random variables.

Conditioning = Projecting!

Chapter 2

Basic Notions of Topology

2.1 Countable and Uncountable Sets

In this section, we investigate the idea of the *cardinality* (or *size*) of a set, with particular emphasis on *countable* sets. We will need these ideas to define *separable* topological spaces, as well as to define the notion of *measure space*, later in this course.

For finite sets, we can determine the size of a set by counting its elements. Thus for example, the set $\{a, b, c\}$ has cardinality 3 (it has 3 elements). We are going to extend this idea of counting to obtain the size of infinite sets, and we will see that infinity comes in many sizes.

First, we explore the idea of *counting*: For the moment, let $\mathbf{n} = \{1, 2, \dots, n\}$ be the set of the first n natural numbers. To say that $A = \{a, b, c\}$ has 3 elements is equivalent to saying that there is a one-to-one correspondence between the sets A and $\mathbf{3}$. Indeed, this is the heart of the idea of counting: When we count the elements of A , we are setting up a bijection between A and $\mathbf{3}$. When we count “One, two, three”, pointing our finger at a, b, c , we are defining a map

$$f : A \rightarrow \mathbf{3} : a \mapsto 1, b \mapsto 2, c \mapsto 3$$

Thus the idea of counting the elements of a finite set X involves finding a bijection between X and some \mathbf{n} . If there is a bijection from X to \mathbf{n} , then X has n elements.

Mathematicians often start counting at zero, i.e. in the mathematical literature, the sets \mathbf{n} are usually defined as

$$\mathbf{n} = \{0, 1, 2, \dots, n - 1\}$$

(We then do not need n to define \mathbf{n} .) This is the convention that we shall adopt henceforth.

It is obvious that two finite sets A and Δ have the same size if and only if there is a one-to-one correspondence $f : A \cong \Delta$. We don't even have to count A and Δ to know that they have the same number of elements. If $A = \{a, b, c, d\}$ and $\Delta = \{\alpha, \beta, \gamma, \delta\}$, then the existence of the bijection $f : A \cong \Delta$ given by

$$f(a) = \beta, f(b) = \delta, f(c) = \alpha, f(d) = \gamma$$

is sufficient to show that A and Δ have the same number of elements. It doesn't tell us that this number is 4. Thus two sets have the same size if and only if there is a bijection between them; we can bypass the idea of number. This is important, because we cannot actually *count* infinite sets. But we can establish bijective correspondences between infinite sets. We shall adopt this idea as our basic idea of size.

Definition 2.1.1 We define an equivalence relation \approx between sets as follows: If A, B are sets, we say that $A \approx B$ if and only if there is a bijection from A to B . If $A \approx B$, we say that A and B have the same **cardinality**. We may also indicate this by saying $|A| = |B|$.

Note that having the same cardinality is an *equivalence relation* between sets, i.e. that

- (i) $|A| = |A|$ (Reflexivity)
- (ii) If $|A| = |B|$, then $|B| = |A|$ (Symmetry)
- (iii) If $|A| = |B|$ and $|B| = |C|$, then $|A| = |C|$ (Transitivity)

Exercise 2.1.2 Prove this assertion. (Note that the assertion is *not obvious*: When we say that $|A| = |B|$, we are not actually claiming that there are two equal numbers. What we *are* saying is that there is a bijection from A to B . To prove (i), for example, you have to find a bijection from A to A .)

□

Examples 2.1.3 (a) Two finite sets have the same cardinality if and only if they have the same number of elements.

- (b) For finite sets, if A is a *proper subset* of B , then $|A| < |B|$. This breaks down completely for infinite sets. Consider, for example, the sets \mathbb{N} and \mathbb{Z} . It is certainly true that $\mathbb{N} \subset \mathbb{Z}$. However, the map $\mathbb{N} \xrightarrow{f} \mathbb{Z}$ defined by

$$f(n) = \begin{cases} \frac{n}{2} & \text{if } n \text{ is even} \\ -\frac{n-1}{2} & \text{if } n \text{ is odd} \end{cases}$$

is a bijection: $f(1) = 0, f(2) = 1, f(3) = -1, f(4) = 2, f(5) = -2, f(6) = 3, \dots$ (Note that we are zig-zagging from the positive integers to the negative integers.) Thus \mathbb{N} and \mathbb{Z} have the same cardinality, even though \mathbb{N} contains fewer elements than \mathbb{Z} .

- (c) We also have $|\mathbb{Q}| = |\mathbb{N}|$. This can be seen as follows. Put the set of strictly positive rational numbers \mathbb{Q}^+ in an array

$$\begin{array}{cccccc} 1/1 & 2/1 & 3/1 & 4/1 & 5/1 & \dots \\ 1/2 & 2/2 & 3/2 & 4/2 & 5/2 & \dots \\ 1/3 & 2/3 & 3/3 & 4/3 & 5/3 & \dots \\ 1/4 & 2/4 & 3/4 & 4/4 & 5/4 & \dots \\ 1/5 & 2/5 & 3/5 & 4/5 & 5/5 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \end{array}$$

We can then trace a zig-zag path that moves through all the rational numbers as follows. Start at the top line and move diagonally down to the left until you reach the leftmost line. Repeat. We thus obtain a sequence

$$\frac{1}{1}, \frac{2}{1}, \frac{1}{2}, \frac{3}{1}, \frac{2}{2}, \frac{1}{3}, \frac{4}{1}, \frac{3}{2}, \frac{2}{3}, \frac{1}{4}, \frac{5}{1}, \dots$$

All of the strictly positive rational numbers occur in this sequence, and they all occur infinitely many times. For example, $\frac{1}{1}, \frac{2}{2}, \frac{3}{3}, \dots$ lie along the diagonal, and they are all equal. To obtain a bijection from \mathbb{N} to \mathbb{Q}^+ , we follow the above sequence of rationals, but we omit any number that has already occurred to ensure that the function is one-to-one, i.e. we *prune* away the repeated values. We therefore define the function $\mathbb{N} \xrightarrow{f} \mathbb{Q}^+$ by

$$f(1) = \frac{1}{1}, f(2) = \frac{2}{1}, f(3) = \frac{1}{2}, f(4) = \frac{3}{1}, f(5) = \frac{1}{3}, f(6) = \frac{4}{1}, \dots$$

Note that $f(5) \neq \frac{2}{2}$, which is after $f(4) = \frac{3}{1}$ in the sequence, because $\frac{2}{2} = \frac{1}{1}$ has already occurred as $f(1)$. Then f is a bijection from \mathbb{N} to \mathbb{Q}^+ . Now even though we haven't found a *formula* for f , it is nevertheless a perfectly good function, and all its values can be calculated. Can you see that $f(16) = \frac{2}{5}$?

In the same way, we can set up a bijection g from \mathbb{N} to the negative rationals. Just put $g(n) = -f(n)$. Finally, we can define a bijection $h : \mathbb{N} \rightarrow \mathbb{Q}$ using f, g and another zig-zag: We define

$$\begin{aligned} h(1) &= 0, h(2) = f(1), h(3) = g(1), h(4) = f(2), \\ h(5) &= g(2), h(6) = f(3), h(7) = g(3), \dots \end{aligned}$$

Again, we have no formula for h , but it is certainly a well-defined function, and all its values can be calculated. Check that $h(23) = -\frac{1}{5}$.

- (d) If A is any set, finite or infinite, then $\mathcal{P}(A) \approx \mathbf{2}^A$. (Recall that $\mathbf{2}^A$ is the set of all functions from A to $\mathbf{2} = \{0, 1\}$). This can be seen as follows: If $B \subseteq A$, define the **indicator function** $I_B : A \rightarrow B$ by

$$I_B(a) = \begin{cases} 1 & \text{if } a \in B \\ 0 & \text{else} \end{cases}$$

Clearly $I_B = I_C$ if and only if $B = C$, and so the map $\mathcal{I} : \mathcal{P}(A) \rightarrow \mathbf{2}^A$ defined by $\mathcal{I}(B) = I_B$ is an injection. Now suppose that $\chi \in \mathbf{2}^A$, i.e. $A \xrightarrow{\chi} \{0, 1\}$. Define a subset $B \subseteq A$ by

$$a \in B \iff \chi(a) = 1$$

It is clear that $\mathcal{I}(B) = I_B = \chi$, and thus that \mathcal{I} is surjective as well. This proves that $|\mathcal{P}(A)| = |\mathbf{2}^A|$.

□

Definition 2.1.4 A set A is said to be *countable* if it is either finite or can be put into a one-to-one correspondence with the natural numbers, i.e. if $|A| = |\mathbf{n}|$ for some $n \in \mathbb{N}$, or $|A| = |\mathbb{N}|$.

Remarks 2.1.5 (a) Basically a set A is countable if its elements can be indexed by the natural numbers, i.e. if it *can* be written as $A = \{a_n : n \in \mathbb{N}\}$. For if A is countable and not finite, then there is a bijection $\mathbb{N} \xrightarrow{f} A$, and we can take $a_n = f(n)$. Conversely, if $A = \{a_n : n \in \mathbb{N}\}$ is infinite, we can define a bijection from \mathbb{N} to A by letting $f(n) = a_n$ (although here some *pruning* is necessary if the a_n aren't all distinct; see Example 2.1.3(c)).

(b) In Examples 2.1.3, we proved that the sets \mathbb{Z} and \mathbb{Q} are countable sets.

(c) The “zig-zag” technique, used above to prove that the rational numbers are countable, is often very useful.

□

Exercise 2.1.6 (a) Show that a set A is countable iff there exists a surjection $f : \mathbb{N} \rightarrow A$.

(b) Show that a set A is countable iff there exists an *injection* $g : A \hookrightarrow \mathbb{N}$.

□

A very basic question that arises is the following: Are all infinite sets countable? As we shall see, the answer is “No!”

Example 2.1.7 We show that the unit interval $I = [0, 1]$ is **uncountable**, i.e. that we cannot find an enumeration

$$I = \{x_n : n \in \mathbb{N}\}$$

The proof is by *contradiction*: Suppose that we *can* find such an enumeration $I = \{x_1, x_2, x_3, x_4, \dots\}$, i.e. that every real number in $[0, 1]$ is equal to x_n for some n . Now every number x_n has a decimal expansion of the form

$$x_n = 0.x_{n1}x_{n2}x_{n3}x_{n4}x_{n5}\dots$$

where x_{nm} is the m^{th} number in the decimal expansion of x_n . Of course some real numbers have two distinct decimal expansions, a terminating one and a non-terminating one. For example, $1.0000\dots = 0.9999\dots$. We will choose the non-terminating decimal expansions for our x_n .

We now create a new real number x from the x_n by a process called *diagonalization*. We choose $a_n \in \{1, 2, \dots, 9\}$ such that the following hold:

$$a_1 \neq x_{11}, a_2 \neq x_{22}, a_3 \neq x_{33}, \dots, a_n \neq x_{nn}, \dots$$

To avoid a situation where we obtain a number x with a terminating decimal expansion, we haven't permitted $a_n = 0$; this is just a technicality. We can now define x : Put

$$x = 0.a_1a_2a_3a_4\dots$$

Here comes the heart of the argument: Clearly $x \in I = [0, 1]$. Now if I can be written as a list $\{x_1, x_2, x_3, \dots\}$, then there must be some n such that $x = x_n$. But the first decimal place of x differs from the first decimal place of x_1 , since $a_1 \neq x_{11}$; hence $x \neq x_1$. Similarly, the second decimal place of x differs from the second decimal place of x_2 , since $a_2 \neq x_{22}$; hence $x \neq x_2$. We can continue in this way to show that $x \neq x_n$ for any $n \in \mathbb{N}$, i.e. x is not on the list $\{x_1, x_2, x_3, \dots\}$.

Given any list x_1, x_2, x_3, \dots of real numbers in $[0, 1]$, this technique allows us to produce real number x that is not on the list. It thus follows that there can be no list containing all the real numbers in $[0, 1]$, i.e. there is no bijection from \mathbb{N} to $[0, 1]$.

□

Hence there are uncountable sets. Clearly \mathbb{R} is also uncountable, because otherwise we could find an enumeration $\{r_1, r_2, r_3, \dots\}$ of \mathbb{R} . By omitting any reals which are not in $[0, 1]$, we could prune this into an enumeration of $[0, 1]$ — and such an enumeration does not exist.

Here is another way of producing uncountable sets:

Exercise 2.1.8 (a) Suppose that A is a set. Show that there is no bijection $A \rightarrow \mathcal{P}(A)$. Conclude that $\mathcal{P}(A)$ is strictly larger than A .

[Hint: Given $f : A \rightarrow \mathcal{P}(A)$, define $B := \{a \in A : a \notin f(a)\}$. Then $B \in \mathcal{P}(A)$, but $B \neq f(a)$ for any $a \in A$.]

(b) Hence show that $\mathcal{P}(\mathbb{N})$ and $2^{\mathbb{N}}$ are uncountable sets.

□

Proposition 2.1.9 (a) If A is countable, and if B is a subset of A , then B is countable.

(b) If A, B are countable, then $A \times B$ is countable.

(c) If A, B are countable, the $A \cup B$ is countable.

(d) If $\mathcal{A} = \{A_n : n \in \mathbb{N}\}$ is a family of countable sets, then $\bigcup_n A_n$ is countable.

Proof: (a) If $\{a_n : n \in \mathbb{N}\}$ is an enumeration of A , we can obtain an enumeration of B by pruning the elements of A which are not in B . This can be accomplished inductively as follows. Let $b_1 = a_n$, where n is the least positive integer such that $a_n \in B$. Suppose now that b_m has been defined and that $b_m = a_i$. Then let $b_{m+1} = a_j$, where j is the least positive integer $> i$ such that $a_j \in B$. Clearly $\{b_m : m \in \mathbb{N}\}$ is an enumeration of B .

(b) One can easily prove that $\mathbb{N} \times \mathbb{N}$ is countable by copying Example 2.1.3(c). Just form an array

$$\begin{array}{ccccccc} (1, 1) & (2, 1) & (3, 1) & (4, 1) & \dots & & \\ 1, 2) & (2, 2) & (3, 2) & (4, 2) & \dots & & \\ (1, 3) & (2, 3) & (3, 3) & (4, 3) & \dots & & \\ \vdots & \vdots & \vdots & \vdots & & & \end{array}$$

and zig-zag your way across this array. Let $A \xrightarrow{f} \mathbb{N}$ and $B \xrightarrow{g} \mathbb{N}$ be bijections. Then the map $h : A \times B \rightarrow \mathbb{N} \times \mathbb{N}$ defined by $h(a, b) = (f(a), g(b))$ is clearly a bijection. Hence $|A \times B| = |\mathbb{N} \times \mathbb{N}| = |\mathbb{N}|$ as required.

(c) follows from (d).

(d) Again we use a zig-zag: Let $\{a_{n1}, a_{n2}, a_{n3}, \dots\}$ be a listing of the elements of A_n . Form an array

$$\begin{array}{cccc} a_{11} & a_{12} & a_{13} & \dots \\ a_{21} & a_{22} & a_{23} & \dots \\ a_{31} & a_{32} & a_{33} & \dots \\ \vdots & \vdots & \vdots & \end{array}$$

and take a path which goes through each element once, pruning duplications.

⊢

Remarks 2.1.10 This proposition shows that you can't make uncountable sets using finite products and countable unions. You *can*, however, make uncountable sets using infinite products and the powerset operation. In Exercise 2.1.8 it is shown that if A is infinite, then $\mathcal{P}(A)$ is uncountable. Similarly if I is infinite, and $|A_i| \geq 2$ for all $i \in I$, then $\prod_I A_i$ is uncountable. We shall not need these facts. Refer to any introductory book on set theory for proofs.

□

2.2 Open Sets and the Interior Operation

The aim of the next two sections is to create a new *language* for talking about *space*. We are going to define a large number of concepts, and prove a large number of simple propositions. We will consider mainly the metric spaces, which include the normed- and inner product spaces as subclasses. To build intuition, you should consider the euclidean spaces, and draw pictures illustrating each concept where possible. All of the propositions in this section are trivial, in that they only require one to plug in the appropriate definitions to prove them. But those definitions take some getting used to. It is therefore *extremely important* that you do all the exercises — perhaps several times over! There is no other way to learn this new language.

Definition 2.2.1 Let (X, d) be a metric space.

- Let $x_0 \in X$ and $r > 0$. The *open ball of radius r centered at x_0* is the set

$$B(x_0, r) := \{x \in X : d(x_0, x) < r\}$$

- Let $A \subseteq X$. A point x_0 is called an *interior point* of A if and only if there is $r > 0$ such that $B(x_0, r) \subseteq A$. In that case, we say that A is a *neighbourhood* of x_0
- If $A \subseteq X$ we define the *interior* of A by

$$A^\circ := \{x \in X : x \text{ is an interior point of } A\}$$

- A subset $U \subseteq X$ is said to be *open* if and only if every point in U is an interior point of U , i.e. iff U is a neighbourhood of all its points.

Exercise 2.2.2 Let (X, d) be \mathbb{R} with the usual metric:

- Show that every open ball is an open interval, and *vice versa*.
- Find an open set which is not an open interval.
- Show that $[0, 1]^\circ = (0, 1)$. Thus $[0, 1]$ is not an open set.
- Show that $\mathbb{Q}^\circ = \emptyset$.

□

Here are some properties of the collection of open sets:

Proposition 2.2.3 Let (X, d) be a metric space.

- Each open ball is an open set.
- A set is open if and only if it is a (possibly infinite) union of open balls.
- The family of open sets satisfies the following axioms:
 - X and \emptyset are open.
 - The union of any (possibly infinite) collection of open sets is open.
 - The intersection any finite collection of open sets is open.

Exercise 2.2.4 Prove Propn. 2.2.3. Here are some hints:

- Given $x \in B(x_0, r)$, you must find an $s > 0$ such that $B(x, s) \subseteq B(x_0, r)$. Explain why you can find a $\varepsilon > 0$ sufficiently small so that $d(x_0, x) + \varepsilon < r$. Then use the Δ -inequality to show that if $y \in B(x, \varepsilon)$, then $d(x_0, y) \leq d(x_0, x) + d(x, y) < r$. Conclude that $B(x, \varepsilon) \subseteq B(x_0, r)$.
- Suppose that A is open. Explain why for each $a \in A$ we can find $r_a > 0$ such that $B(a, r_a) \subseteq A$. Now show $A = \bigcup_{a \in A} B(a, r_a)$.
Conversely, suppose that $A = \bigcup_{i \in I} B(x_i, r_i)$, and that $a \in A$. Then there is $i \in I$ such that $a \in B(x_i, r_i)$. Use (a) to a is conclude that a is an interior point of A .
- (T.1) If A is not open, then there is $x \in A$ such that, for all $r > 0$ we have $B(x, r) \not\subseteq A$. In particular, there is $x \in A$, i.e. A is non-empty.
(T.3) If U_1, \dots, U_n are open, and $x \in U_1 \cap \dots \cap U_n$, then there exist $r_1, \dots, r_n > 0$ such that $B(x, r_i) \subseteq U_i$ for $i = 1, \dots, n$. Take $r = \min\{r_1, \dots, r_n\}$ and explain why $B(x, r) \subseteq U_1 \cap \dots \cap U_n$.

□

Remarks* 2.2.5 A *topological space* is a pair (X, \mathcal{T}) where X is a set and \mathcal{T} is a family of subsets of X with the following properties:

(T.1) $X, \emptyset \in \mathcal{T}$

(T.2) If $U_i \in \mathcal{T}$ for $i \in I$, then $\bigcup_{i \in I} U_i \in \mathcal{T}$.

(T.3) If $U_1, \dots, U_n \in \mathcal{T}$ for some $n \in \mathbb{N}$, then $U_1 \cap \dots \cap U_n \in \mathcal{T}$.

The elements of \mathcal{T} are called the *open sets* of the topological space X . Propn. 2.2.3(c) shows that any metric space is a topological space. The notion of topological space is more general (and more abstract) than that of metric space. In this course, we shall not need this level of generality, however.

□

Exercise* 2.2.6 In the above, the notion of *open set* depends on that of *open ball*, and that, in turn, depends on the notion of *metric*. However, it is possible for two distinct metrics to yield the same family of open sets, as we shall now show. Let d, d_1 be two metrics on \mathbb{R}^n , where d is induced by the usual (euclidean) norm $\|\cdot\|$, and d_1 is induced by the $\|\cdot\|_1$ -norm, given by $\|\mathbf{x}\|_1 := |x_1| + \dots + |x_n|$.

- For the case $n = 2$, draw the open balls $B(\mathbf{0}, 1)$ w.r.t both metrics d, d_1 . Note that the first ball is actually a ball (i.e. round), whereas the second “ball” is diamond-shaped.
- Explain why every d -ball is a d_1 -open set.
- Explain why every d_1 -ball is a d -open set.
- Conclude that every d -open set is a d_1 -open set and vice versa.

The metrics d, d_1 on \mathbb{R}^n are said to be *equivalent*: They generate the same family of open sets, and thus the same topology.

□

Here are some properties of the interior operation:

Proposition 2.2.7 *Let (X, d) be a metric space.*

- $A^\circ \subseteq A$.
- A is open iff $A^\circ = A$.
- $A \subseteq B$ implies $A^\circ \subseteq B^\circ$.
- $A^{\circ\circ} = A^\circ$.
- $(A \cap B)^\circ = A^\circ \cap B^\circ$.
- A° is the union of all open subsets of A .
- A° is the largest open subset of A .

Exercise 2.2.8 (a) Prove Propn. 2.2.7.

- Show that $(A \cup B)^\circ \supseteq A^\circ \cup B^\circ$, and find an example to show that we do not generally have $(A \cup B)^\circ = A^\circ \cup B^\circ$.
- Find an example to show that generally $A^\circ \subseteq B^\circ$ need not imply $A \subseteq B$.

□

2.3 Closed Sets and the Closure Operation

The notion of *closed set* is complementary to that of open set. It is therefore formally superfluous, but many results can be more intuitively comprehended when stated in terms of closed sets, rather than open sets.

Definition 2.3.1 Let (X, d) be a metric space. A subset $C \subseteq X$ is said to be *closed* if and only if its complement $C^c := X - C$ is open.

Exercise 2.3.2 (a) Let (X, d) be \mathbb{R} with the usual metric.

- (i) Show that closed intervals are closed sets.
- (ii) Show that $(0, 1]$ is *neither open, nor closed*.

(b) Use de Morgan's laws to show the analogues of (T.1)–(T.3) for closed sets:

- (i) X and \emptyset are closed.
- (ii) The intersection of a (possibly infinite) collection of closed sets is closed.
- (iii) The union of finitely many closed sets is closed.

(c) Show that the intersection of infinitely many open sets need not be open, and that the union of infinitely many closed sets need not be closed.

□

Exercise 2.3.3 Show that there may be sets which are neither open nor closed. In particular, find a subset of \mathbb{R} (equipped with the usual metric) which is neither open nor closed.

□

We now show that *closed* means *closed under limits*:

Proposition 2.3.4 Suppose that (X, d) is a metric space, and that $C \subseteq X$. Then the following are equivalent:

- (i) C is closed.
- (ii) Whenever $\langle c_n \rangle_n$ is a sequence in C which converges, then it converges to a point in C , i.e.

$$c_n \in C \text{ and } c_n \rightarrow x \quad \text{implies} \quad x \in C$$

Proof: (\Rightarrow) Suppose C is closed in X , and that $\langle c_n \rangle_n$ is a sequence in C which converges, i.e. $c_n \rightarrow x$. We must show $x \in C$, and we argue by *contradiction*: If $x \notin C$, then $x \in C^c$. Since C is closed, C^c is open, so there is $r > 0$ so that $B(x, r) \subseteq C^c$. Since $c_n \rightarrow x$, we have $d(c_n, x) < r$ eventually (i.e. $\exists N \in \mathbb{N} \forall n \geq N [d(c_n, x) < r]$). Then $c_n \in B(x, r) \subseteq C^c$ eventually, contradicting the assumption that $c_n \in C$ for all n .

(\Leftarrow) We prove the *contrapositive*, i.e. we show that $\neg(\text{i})$ implies $\neg(\text{ii})$: Suppose that C is not closed, i.e. that C^c is not open. Then there is a point x in C^c which is not an interior point of C^c . Thus for every $r > 0$ we have $B(x, r) \not\subseteq C^c$. Let $r_n > 0$ be real numbers such that $r_n \rightarrow 0$ (e.g. define $r_n := \frac{1}{n}$). Since $B(x, r_n) \not\subseteq C^c$, we must have $B(x, r_n) \cap C \neq \emptyset$. Choose therefore, $c_n \in B(x, r_n) \cap C$. Then $d(c_n, x) < r_n$, so $c_n \rightarrow x$, yet $x \notin C$.

◄

Definition 2.3.5 Let (X, d) be a metric space, and let $A \subseteq X$.

- A point $x \in X$ is said to be a *cluster point*^a of A iff for every $r > 0$ there exists $y \in B(x, r) \cap A$ such that $y \neq x$.
- A point $x \in X$ is said to be a *boundary point* of A if and only if for every $r > 0$, both

$$B(x, r) \cap A \neq \emptyset \quad \text{and} \quad B(x, r) \cap A^c \neq \emptyset$$

The set of all boundary points of A is called the *boundary* of A , and denoted by ∂A .

- The *closure* of A is defined to be the set A together with all its cluster points, i.e.

$$\bar{A} := \{x \in X : x \in A \text{ or } x \text{ is a cluster point of } A\}$$

^aCluster points are also called *limit points* or *accumulation points* in the literature.

Exercise 2.3.6 (a) Find all the cluster points of the following subsets of \mathbb{R} (equipped with the usual metric):

$$A := (0, 1) \quad B := [0, 1] \quad C := (0, 1) \cap \{2\} \quad D := \{\frac{1}{n} : n \in \mathbb{N}\} \quad E := \mathbb{Q}$$

(b) Find the boundaries $\partial A, \dots, \partial E$ of the sets A, \dots, E above.

(c) Find the closures \bar{A}, \dots, \bar{E} of the sets A, \dots, E above.

□

Remarks 2.3.7 A point x may be a cluster point of a set A without actually belonging to A . The same goes for boundary points. (Of course, if x is an interior point of A , we *must* have $x \in A$).

□

Here are some exercises on cluster points, boundaries and closures:

Exercise 2.3.8 (a) Show that if $A \subseteq B$, then $\bar{A} \subseteq \bar{B}$.

(b) Show that A and A^c have the same boundary, i.e. that $\partial A = \partial A^c$.

(c) Show that x is a cluster point of A iff for every $r > 0$ the set $B(x, r) \cap A$ is an *infinite* set.

[Hints: (c) Suppose that x is a cluster point of A . If $B(x, r) \cap A$ is finite, let $\{y_1, \dots, y_m\}$ be the set of all points in $B(x, r) \cap A$ such that $y_i \neq x$. Now choose s so that $0 < s < \min_{1 \leq i \leq m} d(y_i, x)$ and consider $B(x, s) \cap A$.]

□

Here follows intuitive characterization of open and closed sets in terms of their boundaries:

Proposition 2.3.9 Let (X, d) be a metric space.

(i) $U \subseteq X$ is open iff U contains none of its boundary points, i.e.

$$U \text{ is open} \quad \text{iff} \quad U \cap \partial U = \emptyset$$

(ii) $C \subseteq X$ is closed iff C contains all of its boundary points, i.e.

$$C \text{ is closed} \quad \text{iff} \quad \partial C \subseteq C$$

Proof: (i) (\Rightarrow) : If U is open and $x \in U$, then there is $r > 0$ such that $B(x, r) \subseteq U$. Thus $B(x, r) \cap U^c = \emptyset$, which implies that $x \notin \partial U$. Hence $U \cap \partial U = \emptyset$ when U is open.

(i) (\Leftarrow) : Suppose that $U \cap \partial U = \emptyset$. If $x \in U$, then $x \notin \partial U$. Hence there is $r > 0$ such that $B(x, r) \cap U^c = \emptyset$, from which it follows that $B(x, r) \subseteq U$. Hence x is an interior point of U . Since x was arbitrary, every point of U is an interior point of U , and so U is open.

(ii): We use (i) and the fact that $\partial(C^c) = \partial C$ to argue that

$$\begin{aligned} & C \text{ is closed} \\ \Leftrightarrow & C^c \text{ is open} \\ \Leftrightarrow & C^c \cap \partial(C^c) = \emptyset \\ \Leftrightarrow & \partial C \cap C^c = \emptyset \\ \Leftrightarrow & \partial C \subseteq C \end{aligned}$$

—

The closure of a set can also be characterized in terms of boundary points:

Proposition 2.3.10

$$\bar{A} = A \cup \partial A$$

Proof: First note that

$$\bar{A} - A = \partial A - A$$

i.e. that if $x \notin A$, then x is a boundary point if and only if x is a cluster point of A : For if $x \in \partial A - A$, then each set $B(x, r) \cap A$ contains some point, and that point cannot be x , as $x \notin A$. Thus x is a clusterpoint of A , i.e. $x \in \bar{A} - A$.

Conversely, if $x \in \bar{A} - A$, then $x \in B(x, r) \cap A^c$ for all $r > 0$, and hence both $B(x, r) \cap A^c$ and $B(x, r) \cap A$ are non-empty, as x is a cluster point of A . Hence x is a boundary point of A , i.e. $x \in \partial A - A$.

The result is now obvious: Since $A \subseteq \bar{A}$ we have $\bar{A} = A \cup (\bar{A} - A) = A \cup (\partial A - A) = A \cup \partial A$.

—

The notions of cluster point and closure can seem difficult at first contact, but they're very important. The next few results may result in you pulling out your hair (but it will grow back, so push on).

Proposition 2.3.11 *Let (X, d) be a metric space, and let $A \subseteq X$. Then the following are equivalent:*

- (i) $x \in \bar{A}$
- (ii) There exists a sequence $\langle a_n \rangle_n$ in A such that $a_n \rightarrow x$.
- (iii) For every $r > 0$, we have $B(x, r) \cap A \neq \emptyset$.

Proof: (i) \Rightarrow (ii): Suppose that $x \in \bar{A}$. Then either $x \in A$ or x is a cluster point of A . If x is a cluster point of A is a cluster point of A , choose $a_n \in B(x, \frac{1}{n}) \cap A$. Else, if $x \in A$, define $a_n := x$ for all n . In either case, we have $a_n \in A$ for all n , and $a_n \rightarrow x$.

(ii) \Rightarrow (iii) : Suppose that $\langle a_n \rangle$ is a sequence in A such that $a_n \rightarrow x$, and let $r > 0$ be

arbitrary. Then $a_n \in B(x, r)$ eventually, and so $B(x, r) \cap A \neq \emptyset$.

(iii) \Rightarrow (i): Suppose that for all $r > 0$ we have $B(x, r) \cap A \neq \emptyset$. If $x \in A$, then certainly $x \in \bar{A}$. If $x \notin A$, then (by the definition) x is a cluster point of A , and hence again we conclude that $x \in \bar{A}$.

—

Proposition 2.3.12 *Let (X, d) be a metric space, and let $A \subseteq X$. Then \bar{A} is the smallest closed set which contains A .
In particular, A is closed iff $A = \bar{A}$.*

Exercise 2.3.13 We prove Propn. 2.3.12. Let (X, d) be a metric space, and let $A \subseteq X$.

- (a) We first show, by contradiction, that if $B(x, r) \cap A = \emptyset$ for $x \in X$ and $r > 0$, then the (ostensibly larger) set $B(x, r) \cap \bar{A}$ is empty as well. Assume, therefore, that $B(x, r) \cap A = \emptyset$ but that there exists $y \in B(x, r) \cap \bar{A}$.
 - (a.1) Explain why y is a cluster point of A . [Look at the definition of \bar{A} .]
 - (a.2) Explain why there is $s > 0$ such that $B(y, s) \subseteq B(x, r)$, and conclude that $B(y, s) \cap A = \emptyset$.
 - (a.3) Explain why this is a contradiction.
- (b) Next, we show that \bar{A} is a *closed* set, i.e. that $(\bar{A})^c$ is open.
 - (b.1) Let $x \in (\bar{A})^c$. Use Propn. 2.3.11 to conclude that there is $r > 0$ such that $B(x, r) \cap A = \emptyset$.
 - (b.2) Now use (a) to conclude that $B(x, r) \subseteq (\bar{A})^c$.
 - (b.3) Explain why this is a contradiction, and conclude that \bar{A} is a closed set.
- (c) Finally, we show that \bar{A} is the smallest closed set which contains A , i.e. that if C is a closed set such that $C \supseteq A$, then $C \supseteq \bar{A}$ as well. Assume, therefore, that C is a closed set such that $C \supseteq A$.
 - (c.1) Let $x \in \bar{A}$ be arbitrary. Explain why there is a sequence $\langle a_n \rangle_n$ in A such that $a_n \rightarrow x$.
 - (c.2) Use Propn. 2.3.4 to conclude that $x \in C$.
 - (c.3) Conclude that $\bar{A} \subseteq C$.

□

2.4 Compact Spaces and Sets

Compactness is one of the most important notions in analysis and topology. Yet it is very difficult to explain where the definition comes from. In some ways, compactness as a generalization of *finiteness*. Because the notion is so unfamiliar, we will define two notions of compactness, one in terms of open sets, and one in terms of sequences. We will then show that the two notions coincide in metric spaces¹. It will also transpire that the compact sets in \mathbb{R}^n (with the usual metric) have a simple characterization: They are precisely the closed and bounded sets.

Here is our first definition of compactness — in terms of open sets:

¹They need not coincide in more general topological spaces, though

Definition 2.4.1 Let (X, d) be a metric space, and let $A \subseteq X$

- An open cover of A is a family $\{U_i : i \in I\}$ of open sets in X such that

$$A \subseteq \bigcup_I U_i$$

- The set A is compact if every open cover of A has a finite subcover (i.e. if whenever $\mathcal{U} = \{U_i : i \in I\}$ is an open cover of A , there exists a finite subfamily $U_{i_1}, \dots, U_{i_n} \in \mathcal{U}$ such that $A \subseteq \bigcup_{k=1}^n U_{i_k}$).

Examples 2.4.2 (a) Every finite subset of a metric space (X, d) is compact— why?

- (b) The space \mathbb{R}^n (with the usual metric) is *not* compact. For example, if $U_n = B(0, n)$, then $\{U_n : n \in \mathbb{N}\}$ is an open cover of \mathbb{R}^n . Yet it clearly has no finite subcover — why not? The same argument shows that no unbounded subset of \mathbb{R}^n can be compact, i.e. compact subsets of \mathbb{R}^n are necessarily bounded.
- (c) No open interval (a, b) is compact in \mathbb{R} : Let $U_n = (a + \frac{1}{n}, b - \frac{1}{n})$. Then clearly $(a, b) = \bigcup_n U_n$ (i.e. $\{U_n\}_n$ is an open cover of (a, b)). Yet $\{U_n\}_n$ clearly has no finite subcover of (a, b) — why not?

□

The following exercise shows that there are infinite compact sets in \mathbb{R} :

Exercise 2.4.3 We prove that the closed unit interval $[0, 1]$ is a compact subset of \mathbb{R} .

- (a) Let $I = [0, 1]$ be the closed unit interval, and let $\mathcal{U} = \{U_\gamma : \gamma \in \Gamma\}$ be an open cover of I . Define I^* to be the set of all those $x \in I$ for which $[0, x]$ can be covered by a finite subfamily of \mathcal{U} :

$$I^* = \{x \in [0, 1] : \exists \gamma_1, \dots, \gamma_m \in \Gamma \text{ } ([0, x] \subseteq U_{\gamma_1} \cup \dots \cup U_{\gamma_m})\}$$

- (b) Explain why $0 \in I^*$.
- (c) Show that I^* is a subinterval of I : If $x \in I^*$ and $0 \leq y \leq x$, then $y \in I^*$.
- (d) Define $x^* = \sup I^*$. Explain why $0 \leq x^* \leq 1$.
- (e) Explain why there is $\gamma^* \in \Gamma$ such that $x^* \in U_{\gamma^*}$.
- (f) Explain why $x^* \in I^*$.
- (g) Assume now that $x^* < 1$. Explain why there is $\varepsilon > 0$ such that $[x^* - \varepsilon, x^* + \varepsilon] \subseteq U_{\gamma^*} \cap [0, 1]$.
- (h) Explain why $[0, x^* + \varepsilon]$ can be covered by a finite subfamily of \mathcal{U} .
- (i) Conclude that $x^* + \varepsilon \in I^*$.
- (j) Explain why this is a contradiction.
- (k) Deduce that $1 \in I^*$, and thus that I can be covered by a finite subfamily of \mathcal{U} .

□

The above exercise can easily be generalized to show that any closed interval $[a, b]$ in \mathbb{R} is compact.

We would like to give some other, perhaps more *user friendly*, criteria for *compactness* in metric spaces. We have seen that every compact subset of \mathbb{R} is necessarily bounded. But in \mathbb{R} , the (finite) closed intervals are compact, whereas the open intervals (a, b) are not. *Boundedness* is a requirement, but it is not sufficient:

Exercise 2.4.4 A subset A of a metric space (X, d) is said to be bounded if and only if there exists an open ball $B(x, r)$ such that $A \subseteq B(x, r)$.

- (a) Show that the union of two (and hence of finitely many) bounded sets is bounded.
- (b) Show that a compact set is bounded.
- (c) Show that a bounded set need not be compact.

[Hints: (a) If $A \subseteq B(x_1, r_1)$, $B \subseteq B(x_2, r_2)$ are bounded, show that $A \cup B \subseteq B(x_1, R)$, for any $R \geq \max\{d(x_1, x_2) + r_2, r_1\}$.

(b) Note that if $A \subseteq X$, then, for any $r > 0$, we have $A \subseteq \bigcup_{x \in A} B(x, r)$. If A is compact, then there are $x_1, \dots, x_n \in A$ such that $A \subseteq \bigcup_{k=1}^n B(x_k, r)$. Now use (a).]

□

Proposition 2.4.5 (a) *A compact subset of a metric space is closed and bounded.*

(b) *A closed subset of a compact set is compact.*

Proof: (a) Suppose that K is compact. From Exercise 2.4.4, we know that K is bounded. To prove K is closed, we need only show that its complement K^c is open. Fix $x_0 \in K^c$. For each $y \in K$, let $r_y \in \mathbb{R}$ so that $0 < r_y < \frac{1}{2}d(x_0, y)$. Define

$$U_y := B(x_0, r_y) \quad V_y := B(y, r_y)$$

By construction, we have $x_0 \in U_y$, $y \in V_y$ and $U_y \cap V_y = \emptyset$. Note that $K \subseteq \bigcup_{y \in K} V_y$ (because if $y \in K$, then $y \in V_y$), and thus by compactness there are $y_1, \dots, y_n \in K$ so that $K \subseteq V_{y_1} \cup \dots \cup V_{y_n} := V$. Now define $U := U_{y_1} \cap \dots \cap U_{y_n}$. Then:

- (i) U is open.
- (ii) $x_0 \in U$.
- (iii) $U \cap V = \emptyset$, for if $z \in V$, then $z \in V_{y_j}$ for some $1 \leq j \leq n$, so that $z \notin U_{y_j}$, and hence $z \notin U$.
- (iv) $U \subseteq K^c$, because $U \cap K \subseteq U \cap V = \emptyset$. (Recall that $A \subseteq B^c \Leftrightarrow A \cap B = \emptyset$).

We have thus found an open set U so that $x_0 \in U \subseteq K^c$. Hence x_0 is an interior point of K^c . Since x_0 was arbitrary, all points of K^c are interior points, i.e. K^c is open.

(b) Suppose that K is a compact subset of a metric space (X, d) , and that $C \subseteq K$ is closed. Let $\mathcal{U} := \{U_i : i \in I\}$ be an open cover of C . Since C is closed, $V := C^c$ is open. Now note that if $x \in K$, then either $x \in C$ or $x \in C^c$. In the first case, $x \in \bigcup_I U_i$ (because \mathcal{U} covers C), whereas in the second $x \in V$. In either case $x \in \bigcup_I U_i \cup V$ for any $x \in K$. It follows that $\mathcal{U} \cup \{V\}$ is an open cover of K . As K is compact, there exist $i_1, \dots, i_m \in I$ so that $K \subseteq U_{i_1} \cup \dots \cup U_{i_m} \cup V$. As $V = C^c$, we see that $C \subseteq U_{i_1} \cup \dots \cup U_{i_m}$, i.e. we have found a finite subcover of \mathcal{U} for C .

□

We can say even more:

Proposition 2.4.6 *If a subset K of a metric space is compact, then it is complete, i.e. every Cauchy sequence in K converges, to a limit in K .*

Proof: Suppose that $\langle x_n \rangle_n$ is a Cauchy sequence in K . Then for each $m \in \mathbb{N}$, there exists $N_m \in \mathbb{N}$ so that $d(x_n, x_{N_m}) < \frac{1}{m}$ whenever $n \geq N_m$. Define open subsets

$$G_m := \{y \in X : d(y, x_{N_m}) > \frac{1}{m}\}$$

and note that $x_n \notin G_m$ when $n \geq N_m$. It follows that no finite family of G_m 's covers K : For if $m_1, \dots, m_k \in \mathbb{N}$, we have that $x_n \notin \bigcup_{j=1}^k G_{m_j}$ as soon as $n \geq \max\{N_{m_1}, \dots, N_{m_k}\}$. Because K is compact, it follows that $\bigcup_m G_m \not\supseteq K$, and thus that there is $\bar{x} \in K - \bigcup_m G_m$. We claim that the sequence $\langle x_n \rangle_n$ converges to \bar{x} : If $\varepsilon > 0$, choose $m \in \mathbb{N}$ so that $\frac{2}{m} \leq \varepsilon$. Since $\bar{x} \notin G_m$ we have $d(\bar{x}, x_{N_m}) < \frac{1}{m}$, and thus if $n \geq N_m$, we have

$$d(\bar{x}, x_n) \leq d(\bar{x}, x_{N_m}) + d(x_n, x_{N_m}) < \frac{2}{m} \leq \varepsilon$$

Hence $x_n \rightarrow \bar{x}$, where $\bar{x} \in K$.

—

We are now ready to introduce our second notion of compactness, in terms of sequences:

Definition 2.4.7 A subset K of a metric space (X, d) is said to be sequentially compact if and only if every sequence in K has a subsequence which converges to some element of K .

Example 2.4.8 Any closed and bounded interval $[a, b]$ is sequentially compact in \mathbb{R} . For suppose that $\langle x_n \rangle_n$ is a sequence in $[a, b]$. Then it has a subsequence which converges to $\bar{x} := \limsup_n x_n$. Clearly $\bar{x} \in [a, b]$, and so $\langle x_n \rangle_n$ has a subsequence which converges to a point in $[a, b]$.

□

We are now ready to prove the promised equivalence:

Theorem 2.4.9 Suppose that (X, d) is a metric space and that $K \subseteq X$. The following are equivalent

- (a) K is compact, i.e. every open cover of K has a finite subcover.
- (b) Every infinite subset of K has a cluster point, which belongs to K .
- (c) K is sequentially compact, i.e. every sequence in K has a convergent subsequence whose limit is in K .

Proof: (a) \Rightarrow (b): Suppose that $A \subseteq K$ is infinite, and but that A has no cluster points. Since a set is closed iff it contains all its cluster points, any set without cluster points is necessarily closed. It follows that A is closed, and thus that $V := A^c$ is open. Furthermore, if A has no cluster points, then no $a \in A$ is a cluster point of A , and hence for each $a \in A$ there is an open ball $U_a := B(a, r_a)$ so that $U_a \cap A = \{a\}$. It is now easy to see that $X = \bigcup_{a \in A} U_a \cup V$, and thus that $K \subseteq \bigcup_{a \in A} U_a \cup V$. By compactness of K , there are $a_1, \dots, a_n \in A$ such that $A \subseteq U_{a_1} \cup \dots \cup U_{a_n} \cup V$, and thus that $A \subseteq U_{a_1} \cup \dots \cup U_{a_n}$, as $A \cap V = \emptyset$. But then $A = (U_{a_1} \cap A) \cup \dots \cup (U_{a_n} \cap A) = \{a_1\} \cup \dots \cup \{a_n\} = \{a_1, \dots, a_n\}$, i.e. A is finite — a contradiction. Hence A must have a cluster point.

(b) \Rightarrow (c): Suppose that $\langle x_n \rangle_n$ is a sequence in K . If the set $\{x_n : n \in \mathbb{N}\}$ is finite (i.e. if there are only finitely many different values of the x_n), then $\langle x_n \rangle_n$ obviously has a convergent

subsequence because there must be x such that $x_n = x$ for infinitely many n , and we can take the subsequence to consist of just those x_n — a constant subsequence. Suppose, therefore, that the range $\{x_n : n \in \mathbb{N}\}$ is infinite. Then by (b) it has a clusterpoint $x \in K$, and it is easy to see how to choose a subsequence which converges to x : Simply pick $n_1 < n_2 < \dots$ so that $x_{n_k} \in K \cap B(x, \frac{1}{k})$, for all $k \in \mathbb{N}$.

(c) \Rightarrow (a): Let \mathcal{U} be an open cover of the sequentially compact set K . We first show that with the following property:

$$\exists r > 0 \forall x \in K \exists U \in \mathcal{U} [B(x, r) \subseteq U] \quad (*)$$

If $(*)$ fails, then there is, for each $n \in \mathbb{N}$, a $x_n \in K$ so that $B(x_n, \frac{1}{n}) \not\subseteq U$ for any $U \in \mathcal{U}$. By (c), the sequence $\langle x_n \rangle_n$ has a convergent subsequence $x_{n_k} \rightarrow x$ to some $x \in K$. Now $x \in U$ for some $U \in \mathcal{U}$ (because the U 's cover K and $x \in K$), and hence there is $s > 0$ so that $B(x, s) \subseteq U$. Now choose k sufficiently large that both $d(x_{n_k}, x) < \frac{s}{2}$ and $\frac{1}{n_k} < \frac{s}{2}$. Then

$$y \in B(x_{n_k}, \frac{1}{n_k}) \Rightarrow d(y, x) \leq d(y, x_{n_k}) + d(x_{n_k}, x) < \frac{1}{n_k} + \frac{s}{2} < s \Rightarrow y \in B(x, s)$$

so that $B(x_{n_k}, \frac{1}{n_k}) \subseteq B(x, s) \subseteq U$. This contradicts the definition of x_{n_k} , and hence $(*)$ holds.

With $(*)$ now proved, we can proceed: Choose $r > 0$ so that for every $x \in K$ there is $U \in \mathcal{U}$ so that $B(x, r) \subseteq U$, and let $y_1 \in K$ be arbitrary. If possible, choose inductively $y_{n+1} \in K$ so that $d(y_{n+1}, y_j) \geq r$ for all $j = 1, \dots, n$. If this is possible for all n , one would obtain a sequence $\langle y_n \rangle_n$ in K with $d(y_n, y_m) \geq r$ for all n, m , and such a sequence *cannot* have a convergent subsequence. Hence there is n for which it is impossible to choose y_{n+1} , and hence $K \subseteq \bigcup_{j=1}^n B(y_j, r)$. By definition of r there exist $U_j \in \mathcal{U}$ so that $B(y_j, r) \subseteq U_j$ for $j = 1, \dots, n$. Thus $K \subseteq \bigcup_{j=1}^n U_j$ yields a finite subcover.

—

2.5 Compactness in \mathbb{R}^n

It is easy to characterize compactness in Euclidean space \mathbb{R}^n : The compact sets are precisely those which are closed and bounded. To prove this, we begin with a generalization of the Bolzano–Weierstrass Theorem (cf. Thm A.3.18) to higher (but finite) dimensions:

Theorem 2.5.1 (Bolzano–Weierstrass)

(a) Every bounded sequence in \mathbb{R}^d has a convergent subsequence.

(b) Every infinite bounded subset of \mathbb{R}^d has a cluster point.

Exercise 2.5.2 We prove the Bolzano–Weierstrass Theorem for \mathbb{R}^d . The proof is straightforward, but we need to recall the Bolzano–Weierstrass Theorem for \mathbb{R}^1 : *Every bounded sequence in \mathbb{R} has a convergent subsequence.*

(a) Suppose that $\langle \mathbf{x}_n \rangle_n$ is a bounded sequence in \mathbb{R}^d , where

$$\mathbf{x}_n = (x_n^1, \dots, x_n^d)$$

(Here, x_n^i does *not* mean x_n to the power i — it simply means the i^{th} component of \mathbf{x}_n .) We want to show that $\langle \mathbf{x}_n \rangle_n$ has a convergent subsequence

- (a.1) Explain why there is a subsequence $\langle x_{n'}^1 \rangle_{n'}$ of the sequence $\langle x_n^1 \rangle_n$ which converges to some limit, $x_{n'}^1 \rightarrow a^1$.
- (a.2) Consider now the subsequence $\langle x_{n'}^2 \rangle_{n'}$ of $\langle x_n^2 \rangle_n$, which consists of the same components n' as given in (b.1). Explain why $\langle x_{n'}^2 \rangle_{n'}$ has a further subsequence $\langle x_{n''}^2 \rangle_{n''}$ which converges to some limit, $x_{n''}^2 \rightarrow a^2$. (Again, a^2 is not a -squared, but the second component of a vector \mathbf{a} that we will define below.)
- (a.3) Also explain why also $x_{n''}^1 \rightarrow a^1$.
- (a.4) Next, consider the subsequence $\langle x_{n''}^3 \rangle_{n''}$ of $\langle x_n^3 \rangle_n$ which consists of the same components n'' as in (b.1). Explain why $\langle x_{n''}^3 \rangle_{n''}$ has a further subsequence $\langle x_{n'''}^3 \rangle_{n'''}$ which converges to some limit, $x_{n'''}^3 \rightarrow a^3$.
- (a.5) Also explain why also $x_{n'''}^1 \rightarrow a^1$ and $x_{n'''}^2 \rightarrow a^2$.
- (a.6) We can proceed in this way, component by component, until we produce a subsequence $\langle x_{n^{(d)}}^d \rangle_{n^{(d)}}$ of $\langle x_n^d \rangle_n$ with the properties that

$$x_{n^{(d)}}^1 \rightarrow a^1, \quad x_{n^{(d)}}^2 \rightarrow a^2, \quad \dots \quad x_{n^{(d)}}^d \rightarrow a^d$$

Thus we have $\mathbf{x}_{n^{(d)}} \rightarrow \mathbf{a}$ in \mathbb{R}^d , where $\mathbf{a} := (a^1, a^2, \dots, a^n)$.

- (b) Suppose that $A \subseteq \mathbb{R}^d$ is a bounded infinite set. We must show that A has a cluster point.
- b.1 First explain why we may choose a sequence $\langle \mathbf{x}_n \rangle_n$ in A so that $\mathbf{x}_n \neq \mathbf{x}_m$ when $n \neq m$.
- b.2 By (a), there is a subsequence $\langle \mathbf{x}_{n'} \rangle_{n'}$ which converges. Explain why $\mathbf{x} := \lim_n \mathbf{x}_n$ is a cluster point of A .

□

Theorem 2.5.3 (Heine-Borel) *A subset of \mathbb{R}^d is compact if and only if it is closed and bounded.*

Proof: (\Rightarrow): This follows immediately from Propn. 2.4.5(b).

(\Leftarrow): Suppose that $C \subseteq \mathbb{R}^d$ is closed and bounded, and let $\langle \mathbf{x}_n \rangle_n$ be a sequence in C . By the Bolzano–Weierstrass Theorem, $\langle \mathbf{x}_n \rangle_n$ has a convergent subsequence, $\mathbf{x}_{n'} \rightarrow \mathbf{x}$. Because C is closed, we have $\mathbf{x} \in C$. Hence every sequence in C has a subsequence which converges to a point in C , i.e. C is sequentially compact.

⊢

2.6 Convergence and Continuity

In this section we take another look at continuity. To do that, we need a little set theory:

2.6.1 Pull-backs and Push-forwards

Definition 2.6.1 Let X, Y be sets, and let $f : X \rightarrow Y$ be an arbitrary function.

- If $A \subseteq X$, then the *push-forward* (or *direct image*) of A along f is the set $f[A] \subseteq Y$ defined by

$$f[A] := \{f(x) : x \in A\} = \{y \in Y : \exists x \in A (y = f(x))\}$$

- If $B \subseteq Y$ then the *pull-back* (or *inverse image*) of B along f is the set $f^{-1}[B] \subseteq X$ defined by

$$f^{-1}[B] := \{x \in X : f(x) \in B\}$$

Thus

$$x \in f^{-1}[B] \quad \Longleftrightarrow \quad f(x) \in B$$

Remarks 2.6.2 IMPORTANT: The definition of the inverse image does *not* require the existence of an inverse function, i.e. it is not necessary that the function f^{-1} exists² in the definition of inverse image. Observe that the definition of $f^{-1}[B]$, namely $\{x \in X : f(x) \in B\}$ involves only the function f (and does not involve f^{-1}) and is therefore well-defined if f is.

When $f : X \rightarrow Y$ is invertible, however, then the inverse image of $B \subseteq Y$ along f and the direct image of B along $f^{-1} : Y \rightarrow X$ are easily seen to coincide.

□

Exercises 2.6.3 (a) Let $X := \{0, 1, 2, 3, 4\}$, $Y := \{a, b, c, d\}$ and define $f : X \rightarrow Y$ by

$$0 \mapsto a, \quad 1 \mapsto d, \quad 2 \mapsto c, \quad 3 \mapsto a, \quad 4 \mapsto c$$

Determine the direct images of the following sets:

$$A_1 := \{0\} \quad A_2 := \{0, 1\} \quad A_3 := X \quad A_4 = \emptyset$$

Determine the inverse images of the following sets:

$$B_1 := \{a\} \quad B_2 = \{b\} \quad B_3 = \{a, b, c\}, \quad B_4 := \emptyset$$

(b) Let $g : \mathbb{R}^2 \rightarrow \mathbb{R} : (x, y) \mapsto x^2 + y^2$. Determine

$$g^{-1}\{1\} \quad g^{-1}[0, 1] \quad g^{-1}(-\infty, 0)$$

(c) Let $h : \mathbb{R} \rightarrow \mathbb{R} : x \mapsto x^2$. Determine

$$h^{-1}[1, 2] \quad h[1, 2]$$

□

Note that if $f : X \rightarrow Y$, then $f[\cdot]$ is a function which assigns to every subset of X a subset of Y , i.e.

$$f[\cdot] : \mathcal{P}(X) \rightarrow \mathcal{P}(Y) : A \subseteq X \mapsto f[A] \subseteq Y$$

Similarly, $f^{-1}[\cdot]$ is a function which assigns to every subset of Y a subset of X , i.e.

$$f^{-1}[\cdot] : \mathcal{P}(Y) \rightarrow \mathcal{P}(X) : B \subseteq Y \mapsto f^{-1}[B] \subseteq X$$

Lemma 2.6.4 If $f : X \rightarrow Y$, $A \subseteq X$ and $B \subseteq Y$, then

$$f[A] \subseteq B \quad \text{iff} \quad A \subseteq f^{-1}[B]$$

Proof: (\Rightarrow): Suppose that $f[A] \subseteq B$. If $x \in A$, then $f(x) \in f[A]$. Hence $f(x) \in B$, and thus $x \in f^{-1}[B]$. Thus $x \in A$ implies $x \in f^{-1}[B]$.

(\Leftarrow): Suppose that $A \subseteq f^{-1}[B]$. If $y \in f[A]$, then there is $x \in A$ such that $f(x) = y$. But $x \in A$ implies $f(x) \in B$, and hence $y \in B$. Thus $y \in f[A]$ implies $y \in B$.

◄

The pull-back function $f^{-1}[\cdot]$ is particularly well-behaved with respect to the algebraic set operations. :

²Recall that f^{-1} exists iff f is a *bijection* i.e. both *one-to-one* and *onto*.

Proposition 2.6.5 Suppose that $f : X \rightarrow Y$ is a function, and that $B_i \subseteq Y$ for $i \in I$.

- (a) $f^{-1}[\bigcup_{i \in I} B_i] = \bigcup_{i \in I} f^{-1}[B_i]$.
- (b) $f^{-1}[\bigcap_{i \in I} B_i] = \bigcap_{i \in I} f^{-1}[B_i]$.
- (c) $f^{-1}[B]^c = f^{-1}[B^c]$. (Here $f^{-1}[B]^c$ is the complement of $f^{-1}[B]$ in X , whereas B^c is the complement of B in Y .)

Proof: (a)

$$\begin{aligned}
 & x \in f^{-1} \left[\bigcup_i B_i \right] \\
 \text{iff} \quad & f(x) \in \bigcup_i B_i \\
 \text{iff} \quad & (\exists i \in I)(f(x) \in B_i) \\
 \text{iff} \quad & (\exists i \in I)(x \in f^{-1}[B_i]) \\
 \text{iff} \quad & x \in \bigcup_i f^{-1}[B_i]
 \end{aligned}$$

+

Exercise 2.6.6 (a) Prove the remainder of Propn. 2.6.5.

(b) Investigate the analogous results for the push-forward function $f[\cdot]$.

□

2.6.2 Topological Characterizations of Convergence and Continuity

We are now able to characterize notions related to convergence and continuity purely in terms of neighbourhoods and open sets. First, we show how to define the notion of *convergence*:

Proposition 2.6.7 Let $\langle x_n \rangle_n$ be a sequence in a metric space (X, d) and let $x \in X$. Then the following are equivalent:

- (i) $x_n \rightarrow x$.
- (ii) For any neighbourhood A of x , we have $x_n \in A$ eventually.

Proof: (i) \Rightarrow (ii): Suppose that $x_n \rightarrow x$, and that A is a neighbourhood of x . Choose $r > 0$ so that $B(x, r) \subseteq A$, and then choose N such that $d(x_n, x) < r$ whenever $n \geq N$. Then $x_n \in A$ whenever $n \geq N$.

(ii) \Rightarrow (i): Let $\varepsilon > 0$. Then $B(x, \varepsilon)$ is a neighbourhood of x . Hence there exists $N \in \mathbb{N}$ such that $x_n \in B(x, \varepsilon)$ whenever $n \geq N$. Thus $d(x_n, x) < \varepsilon$ whenever $n \geq N$.

+

Theorem 2.6.8 Let (X, d_X) and (Y, d_Y) be metric spaces, and let $f : X \rightarrow Y$.

- (a) f is continuous at $x_0 \in X$ iff for every (open) neighbourhood V of $f(x_0)$ there exists an (open) neighbourhood U of x_0 such that $f[U] \subseteq V$.
- (b) f is everywhere continuous iff whenever V is an open subset of Y , then $f^{-1}[V]$ is an open subset of X .

Proof: (a): Suppose that $f : X \rightarrow Y$ is continuous at the point $x_0 \in X$, and let V be a neighbourhood of $f(x_0) \in Y$. By definition of *neighbourhood* there is $\varepsilon > 0$ such that $B_Y(f(x_0), \varepsilon) \subseteq V$. Because f is continuous, there is $\delta > 0$ such that $d_X(x_0, x) < \delta$ implies $d_Y(f(x_0), f(x)) < \varepsilon$. Thus $U := B_X(x_0, \delta)$ is an open neighbourhood of x_0 with the property that $f[U] \subseteq V$.

Conversely, suppose that for every (open) neighbourhood V of $f(x_0)$ there exists an (open) neighbourhood U of x_0 such that $f[U] \subseteq V$. To prove that f is continuous at x_0 it suffices to show that whenever $x_n \rightarrow x_0$, we must have $f(x_n) \rightarrow f(x_0)$. So suppose that $x_n \rightarrow x_0$ in X . Let V be an arbitrary neighbourhood of $f(x_0)$, and choose a neighbourhood U of x_0 such that $f[U] \subseteq V$. Since $x_n \in U$ eventually by Propn. 2.6.7, we must have $f(x_n) \in V$ eventually. It follows from Propn. 2.6.7 that $f(x_n) \rightarrow f(x_0)$.

(b): Suppose that f is continuous, and that $V \subseteq Y$ is open. If $x_0 \in f^{-1}[V]$, then V is an open neighbourhood of $f(x_0)$, and so there is an open neighbourhood U of x_0 such that $f[U] \subseteq V$. Then $x_0 \in U \subseteq f^{-1}[V]$, and hence x_0 is an interior point of $f^{-1}[V]$. Since $x_0 \in f^{-1}[V]$ was arbitrary, every point of $f^{-1}[V]$ is an interior point, i.e. $f^{-1}[V]$ is open.

Conversely, suppose that pullbacks of open sets are open. If $x_0 \in X$, and V is an open neighbourhood of $f(x_0)$, then $U := f^{-1}[V]$ is an open neighbourhood of x_0 with the property that $f[U] \subseteq V$. By (a), f is continuous at any $x_0 \in X$.

◄

2.6.3 Continuity and Compactness

The following result has many important applications:

Theorem 2.6.9 *If $f : (X, d_X) \rightarrow (Y, d_Y)$ is continuous and K is a compact subset of X , then $f[K]$ is a compact subset of Y .*

We will give two proofs, one in terms of sequences, and one in terms of open covers:

Exercise 2.6.10 Use the equivalence of *compactness* and *sequential compactness* to give a proof of Thm. 2.6.9, i.e. prove that if f is continuous and K sequentially compact, then $f[K]$ is sequentially compact.

□

Alternate Proof of Thm. 2.6.9: Let $\mathcal{V} := \{V_i : i \in I\}$ be an open cover of $f[K]$ in Y . We show that there exist finitely many $i_1, \dots, i_m \in I$ such that $f[K] \subseteq V_{i_1} \cup \dots \cup V_{i_m}$. Define $U_i := f^{-1}[V_i]$ for $i \in I$, and observe that each U_i is open, by Propn. 2.6.8. Moreover

$$x \in K \Rightarrow f(x) \in f[K] \Rightarrow \exists i \in I (f(x) \in V_i) \Rightarrow \exists i \in I (x \in f^{-1}[V_i]) \Rightarrow x \in \bigcup_{i \in I} U_i$$

Hence $\mathcal{U} := \{U_i : i \in I\}$ is an open cover of K . As K is compact, there exist $i_1, \dots, i_m \in I$ such that $K \subseteq U_{i_1} \cup \dots \cup U_{i_m}$. Now $f[K] \subseteq f[\bigcup_{j=1}^m U_{i_j}] = f[f^{-1}[\bigcup_{j=1}^m V_{i_j}]] \subseteq \bigcup_{j=1}^m V_{i_j}$.

◄

The above theorem has the following important consequences for *optimization*. It states that a continuous function possesses a maximum and minimum on any compact set:

Corollary 2.6.11 Suppose that $f : (X, d_X) \rightarrow \mathbb{R}$ is continuous, and that K is a compact subset of X . Then f attains its supremum and infimum on K , i.e. there exist $k^*, k_* \in K$ such that

$$f(k^*) = \sup f[K] = \sup_{k \in K} f(k) \quad f(k_*) = \inf f[K] = \inf_{k \in K} f(k)$$

Proof: $C := f[K]$ is compact, and hence closed and bounded. From the fact that C is bounded, it follows that $\sup C$ and $\inf C$ exist and are finite (by the Completeness Axiom). Since C is closed, we must have $\sup C, \inf C \in C$. (To see, e.g., that $c^* := \sup C \in C$, note that for every $n \in \mathbb{N}$ there is $c_n \in C$ such that $c^* - \frac{1}{n} < c_n \leq c^*$, so that $\langle c_n \rangle_n$ is a sequence in C with $c_n \rightarrow c^*$. Closedness of C now guarantees that $c^* \in C$.) In particular, $\sup f[K] \in f[K]$, i.e. there is $k^* \in K$ so that $\sup f[K] = f(k^*)$. The same holds for $\inf f[K]$.

—

Finally, we show that continuous functions on compact sets possess a stronger property, that of *uniform continuity*:

Definition 2.6.12 Let $(X, d_X), (Y, d_Y)$ be metric spaces and let $A \subseteq X$. A function $f : A \rightarrow Y$ is said to be *uniformly continuous* on A if and only if

$$\forall \varepsilon > 0 \exists \delta > 0 \forall x, y \in A [d_X(x, y) < \delta \rightarrow d_Y(f(x), f(y)) < \varepsilon]$$

Remarks 2.6.13 Note the difference between the definition of *continuity* and *uniform continuity*: $f : (X, d_X) \rightarrow (Y, d_Y)$ is *continuous* iff it is continuous at every $x \in X$, i.e. iff

$$\forall x \in X \forall \varepsilon > 0 \exists \delta > 0 \forall y \in X [d_X(x, y) < \delta \rightarrow d_Y(f(x), f(y)) < \varepsilon]$$

Thus given an $x \in X$ and a ε , we can find a δ , and that δ may depend on both ε and x .

On the other hand, $f : (X, d_X) \rightarrow (Y, d_Y)$ is *uniformly continuous* iff

$$\forall \varepsilon > 0 \exists \delta > 0 \forall x, y \in X [d_X(x, y) < \delta \rightarrow d_Y(f(x), f(y)) < \varepsilon]$$

Thus given a ε we can find a δ , and that δ will work for *uniformly for all x* . It does not make sense to talk about uniform continuity at a point — we are discussing all x in a given domain at once.

□

Exercise 2.6.14 (a) Show that if f is uniformly continuous, then it is continuous at any point.

(b) Show that any affine function $f : \mathbb{R} \rightarrow \mathbb{R} : x \mapsto ax + b$ is uniformly continuous on all of \mathbb{R} .

(c) Show that the function $g : \mathbb{R} \rightarrow \mathbb{R} : x \mapsto x^2$ is not uniformly continuous on \mathbb{R} , but that it is uniformly continuous on $[0, 1]$ (or, indeed, any bounded subset of \mathbb{R}).

(d) Show that the function $h : (0, \infty) \rightarrow \mathbb{R} : x \mapsto \frac{1}{x}$ is uniformly continuous on (r, ∞) for any $r > 0$, but not uniformly continuous on $(0, \infty)$.

[Hints: (c) Given $\delta > 0$, note that $|(x + \delta)^2 - x^2| \geq 2\delta|x|$ can be made arbitrarily large by choosing x sufficiently large. Conclude that, with $\varepsilon = 1$ (or any other value > 0), there is no $\delta > 0$ such that $|g(y) - g(x)| < \varepsilon$ for all x, y with $|x - y| < \delta$.]

□

2.7 Separable Spaces

Intuitively, a subset D of a metric space (X, d) is *dense* when it is “spread everywhere” within the space X . This does not mean that $D = X$. For example, the set \mathbb{Q} of rational numbers is spread every where within the space \mathbb{R} of reals: Between any two real numbers one can find a rational number. As we have seen, \mathbb{R} is much bigger than \mathbb{Q} , i.e. \mathbb{Q} is quite “small” — \mathbb{Q} is countable, \mathbb{R} is not. The fact that every real number can be approximated arbitrarily closely by rational numbers can be very useful, as we need deal only with a much smaller set. It deserves to be formalized:

Definition 2.7.1 Let (X, d) be a metric space. A subset $D \subseteq X$ is said to be *dense* in X if

$$D \cap U \neq \emptyset \quad \text{for every non-empty open set } U$$

A metric space is said to be *separable* if it has a countable dense subset.

Examples 2.7.2 (a) \mathbb{R}^d is separable, as $\mathbb{Q}^d = \mathbb{Q} \times \cdots \times \mathbb{Q}$ is countable, by Propn. 2.1.9.

(b) Every compact space is separable: For suppose (X, d) is compact. Then, for every $n \in \mathbb{N}$, there exist finitely many open balls of radius $\frac{1}{n}$ which cover X . Let D_n be the set of centers of those finitely many open balls, and let $D := \bigcup_n D_n$. Then D is a countable union of finite sets, and hence countable, by Propn. 2.1.9. We claim that D is dense. For suppose that $U \subseteq X$ is non-empty and open. Let $x \in U$, and choose $n \in \mathbb{N}$ so that $B(x, \frac{1}{n}) \subseteq U$. As $\bigcup_{y \in D_n} B(y, \frac{1}{n})$ covers X , there is $d \in D_n$ such that $x \in B(d, \frac{1}{n})$. Clearly, then $d \in B(x, \frac{1}{n})$, so that $d \in D \cap B(x, \frac{1}{n}) \subseteq D \cap U$. Hence $D \cap U \neq \emptyset$. □

2.8 The Banach Space $\mathcal{C}[a, b]$

We investigate, in this section, the space $(\mathcal{C}[a, b], \|\cdot\|_\infty)$ of all continuous functions $f : [a, b] \rightarrow \mathbb{R}$, equipped with the sup-norm: $\|f\|_\infty := \sup_{x \in [a, b]} |f(x)|$. We will first investigate the notion of convergence in this space — uniform convergence — and show that it is a Banach space (i.e. a complete normed space). Then we give a criterion for a subset of this space to be compact — the Arzelà–Ascoli Theorem. Finally, we show that $\mathcal{C}[a, b]$ is separable: The set of polynomials with rational coefficients is dense in $\mathcal{C}[a, b]$.

2.8.1 Uniform Convergence

Let (X, d) be a metric space, and let $\mathcal{C}(X)$ be the set of all continuous functions $f : X \rightarrow \mathbb{R}$. We will be interested mainly in the case where $X = [a, b]$ is a compact interval in \mathbb{R} .

$\mathcal{C}(X)$ inherits a lot of structure from \mathbb{R} . It is easy to see that

- $\mathcal{C}(X)$ is a linear space, when addition and scalar multiplication are defined pointwise, i.e.

$$(f + g)(x) := f(x) + g(x) \quad (\alpha f)(x) := \alpha \cdot f(x) \quad \text{for } x \in X, \alpha \in \mathbb{R}$$

- When X is compact, the function $\|\cdot\|_\infty : \mathcal{C}(X) \rightarrow \mathbb{R}$ defined by

$$\|f\|_\infty := \sup_{x \in X} |f(x)|$$

is a norm on $\mathcal{C}(X)$. This norm is called the *uniform* norm or the *sup* norm.

[Note that $\|f\|_\infty < \infty$ when f is continuous, because X is compact — cf. Corollary 2.6.11.]

The elements of $\mathcal{C}(X)$ are functions, so two notions of convergence suggest themselves:

Definition 2.8.1 Let $f_n, f \in \mathcal{C}(X)$ for $n \in \mathbb{N}$.

- (a) We say that $f_n \rightarrow f$ *pointwise* if and only if $f_n(x) \rightarrow f(x)$ for all $x \in X$.
- (b) We say that $f_n \rightarrow f$ *uniformly* if and only if it converges w.r.t. the uniform norm, i.e. iff $\|f_n - f\|_\infty \rightarrow 0$.

□

Examples 2.8.2 (a) Suppose that $f_n : [0, 1] \rightarrow \mathbb{R} : x \mapsto x(x + \frac{1}{n})$. Then as $n \rightarrow \infty$, we have $f_n(x) \rightarrow x^2$. So if $f(x) := x^2$, then $f_n \rightarrow f$ pointwise on $[0, 1]$.

We also have $\|f_n - f\|_\infty = \sup_{0 \leq x \leq 1} |x| \frac{1}{n} = \frac{1}{n}$ so that $\|f_n - f\|_\infty \rightarrow 0$. Hence we also have $f_n \rightarrow f$ uniformly.

(b) Let $f_n : [0, 1] \rightarrow \mathbb{R} : x \mapsto x^n$. Then

$$\lim_n f_n(x) = \begin{cases} 0 & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x = 1 \end{cases}$$

Hence

$$f_n \rightarrow f \text{ pointwise on } [0, 1]$$

where $f : [0, 1] \rightarrow \mathbb{R}$ is defined by

$$f(x) = \begin{cases} 0 & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x = 1 \end{cases}$$

On the other hand

$$\|f_n - f\|_\infty = \sup_{0 \leq x < 1} |x^n| = 1 \quad \text{for all } n$$

Hence $f_n \not\rightarrow f$ uniformly.

□

Exercise 2.8.3 For $n \in \mathbb{N}$, define $f_n : [0, 1] \rightarrow \mathbb{R}$ by

$$f_n(x) = \frac{x^n}{1 + x^n}$$

- (a) Show that f_n converges pointwise, and determine the limit function $f = \lim_n f_n$.
- (b) Does $f_n \rightarrow f$ uniformly?

□

How do pointwise and uniform convergence differ? Look at the ε -forms of the definitions:

$$\begin{aligned} f_n \rightarrow f \text{ pointwise} & \quad \text{iff} \quad (\forall x \in X)(\forall \varepsilon > 0)(\exists N \in \mathbb{N})(\forall n \in \mathbb{N})[n \geq N \Rightarrow |f_n(x) - f(x)| < \varepsilon] \\ f_n \rightarrow f \text{ uniformly} & \quad \text{iff} \quad (\forall \varepsilon > 0)(\exists N \in \mathbb{N})((\forall n \in \mathbb{N})\forall x \in X)[n \geq N \Rightarrow |f_n(x) - f(x)| < \varepsilon] \end{aligned}$$

The difference between these notions of convergence is clear: The order of the quantifiers is different: In the case of pointwise convergence, for each ε and each x , we can find an N such that... That N may depend on ε as well as x .

In the case of uniform convergence, for each ε we can find an N such that... That N may depend on ε , but not on a particular x : We can find an N which works for all x simultaneously.

Remarks 2.8.4 When X is a compact subinterval of \mathbb{R} , the concept of uniform convergence has a neat graphical interpretation: $f_n \rightarrow f$ uniformly if and only if the graphs of the f_n lie eventually within a band of radius ε about the graph of f , i.e. if and only if at most finitely many of the f_n have some portion of their graphs lying outside this band.

□

Proposition 2.8.5 *If $f_n \rightarrow f$ uniformly, then $f_n \rightarrow f$ pointwise.*

Exercise 2.8.6 Prove Propn. 2.8.5.

□

In Examples 2.8.2 and Exercise 2.8.3 we saw that the pointwise limit of continuous functions need not be continuous. This “pathology” cannot happen with uniform convergence:

Theorem 2.8.7 *Suppose that $f_n \rightarrow f$ uniformly on X . If each f_n is continuous at $x_0 \in X$, then f is also continuous at x_0 .*

Proof: Let $\varepsilon > 0$. We must find a $\delta > 0$ such that $|f(x) - f(x_0)| < \varepsilon$ whenever $d(x, x_0) < \delta$ (for $x \in X$).

First, by uniform convergence, we may choose N such that $|f_n(x) - f(x)| < \frac{\varepsilon}{3}$ for all $x \in X$, whenever $n \geq N$.

Then, because f_N is continuous at x_0 , we may choose $\delta > 0$ such that $|f_N(x) - f_N(x_0)| < \frac{\varepsilon}{3}$ whenever $d(x, x_0) < \delta$.

It follows that

$$|f(x) - f(x_0)| \leq |f(x) - f_N(x)| + |f_N(x) - f_N(x_0)| + |f_N(x_0) - f(x_0)| < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3}$$

whenever $d(x, x_0) < \delta$.

□

Corollary 2.8.8 *The uniform limit of continuous functions is a continuous function.*

Theorem 2.8.9 (Cauchy criterion for uniform convergence)

Suppose that $\langle f_n \rangle_n$ is a sequence of functions with common domain X . There exists a function f such that $f_n \rightarrow f$ uniformly on X if and only if

$$(\forall \varepsilon > 0)(\exists N \in \mathbb{N})(\forall n, m > N)(\forall x \in X)[|f_n(x) - f_m(x)| < \varepsilon]$$

i.e. for every $\varepsilon > 0$ we can find an N such that whenever $n, m > N$ and $x \in X$, we have $|f_n(x) - f_m(x)| < \varepsilon$.

Exercise 2.8.10 Prove Theorem 2.8.9.

□

We can now prove the following theorem:

Theorem 2.8.11 *If (X, d) is a compact metric space, then $(\mathcal{C}(X), \|\cdot\|_\infty)$ is a Banach space.*

Proof: We already know that $(\mathcal{C}(X), \|\cdot\|_\infty)$ is a normed space, so it remains to check that it is complete (w.r.t. the sup norm).

If $\langle f_n \rangle_n$ is a Cauchy sequence in $\mathcal{C}(X)$, then for all $x \in X$, the component sequence $\langle f_n(x) \rangle_n$ is a Cauchy sequence in \mathbb{R} . As \mathbb{R} is complete, each component sequence converges to some real number—call it $f(x)$. This convergence is also uniform: If $x \in X$, then $|f_n(x) - f(x)| = \lim_m |f_n(x) - f_m(x)| \leq \limsup_m \|f_n - f_m\|_\infty$, and hence $\|f_n - f\|_\infty \leq \limsup_m \|f_n - f_m\|_\infty$. Since $\langle f_n \rangle_n$ is a Cauchy sequence in $\mathcal{C}(X)$, we have $\lim_{m,n \rightarrow \infty} \|f_n - f_m\|_\infty \rightarrow 0$, and so $\|f_n - f\|_\infty \rightarrow 0$ also.

Since the uniform limit of continuous functions is again continuous, we see that $f \in \mathcal{C}(X)$, and thus that $\mathcal{C}(X)$ is complete when equipped with the sup norm.

—

Exercise 2.8.12 (a) Show that $\mathcal{C}[0, 1]$ is an *infinite dimensional* space.

[Hint: Consider the functions $f_n(x) := x^n$.]

(b) Show that the closed unit ball $\bar{B}(0, 1) := \{f \in \mathcal{C}[0, 1] : \|f\|_\infty \leq 1\}$ is a closed and bounded set which is not compact.

[Hint: The functions $f_n(x) := x^n$ form a sequence in $\bar{B}[0, 1]$ which has no convergent subsequence.]

□

Lemma 2.8.13 (Dini's Theorem) *If $f_n, f \in \mathcal{C}(K)$ and $f_n \downarrow f$, then $f_n \rightarrow f$ uniformly.*

Proof: Define $G_n := \{x \in K : f_n(x) - f(x) < \varepsilon\}$. Each G_n is open, $G_1 \subseteq G_2 \subseteq G_3 \subseteq \dots$ and $\bigcup_n G_n = K$. Since K is compact, there is n_0 such that $G_{n_0} = K$. Then $|f_n(x) - f(x)| < \varepsilon$ for all $n \geq n_0$ and all $x \in K$.

—

Definition 2.8.14 (i) A subset $\mathcal{F} \subseteq \mathcal{C}(K)$ is said to be *equicontinuous* at a point $x_0 \in K$ if and only if $\sup_{f \in \mathcal{F}} |f(x) - f(x_0)| \rightarrow 0$ as $x \rightarrow x_0$, i.e. iff

$$\forall \varepsilon > 0 \exists \delta > 0 \forall f \in \mathcal{F} \forall x \in K [d(x, x_0) < \delta \implies |f(x) - f(x_0)| < \varepsilon]$$

(This is just the usual definition of continuity of f at x_0 , but here we require that for a given $\varepsilon > 0$ there is a δ which “works” for all $f \in \mathcal{F}$ simultaneously.)

(ii) $\mathcal{F} \subseteq \mathcal{C}(K)$ is said to be equicontinuous iff it is equicontinuous at every $x_0 \in K$.

(iii) $\mathcal{F} \subseteq \mathcal{C}(K)$ is said to be *uniformly equicontinuous* iff

$$\forall \varepsilon > 0 \exists \delta > 0 \forall f \in \mathcal{F} \forall x, y \in K [d(x, y) < \delta \implies |f(x) - f(y)| < \varepsilon]$$

(This is just the usual definition of uniform continuity of f , but here we require that for a given $\varepsilon > 0$ there is a δ which “works” for all $f \in \mathcal{F}$ simultaneously.)

The following lemma extends the result that continuous functions are uniformly continuous on compact sets:

Lemma 2.8.15 *If (K, d) is a compact metric space, then any equicontinuous family is uniformly equicontinuous.*

Proof: Suppose $\mathcal{F} \subseteq \mathcal{C}(K)$ is equicontinuous, but not uniformly equicontinuous. Then there exist $\varepsilon > 0$, $x_n, y_n \in K$ and $f_n \in \mathcal{F}$ such that for all $n \in \mathbb{N}$ we have

$$d(x_n, y_n) < \frac{1}{n}, \quad \text{yet} \quad |f_n(x_n) - f_n(y_n)| > \varepsilon$$

Since K is compact, the sequence x_n has a convergent subsequence $x_{n_k} \rightarrow x$, so we may w.l.o.g. assume $x_n \rightarrow x$. Then also $y_n \rightarrow x$. As \mathcal{F} is equicontinuous at x , we have, for sufficiently large n , that

$$|f_n(y_n) - f_n(x)| < \frac{\varepsilon}{2} \quad \text{and} \quad |f_n(x_n) - f_n(x)| < \frac{\varepsilon}{2}$$

from which we obtain the contradiction $|f_n(x) - f_n(y)| < \varepsilon$.

—

A related notion, which may help to clarify the above, is the *modulus of continuity*. This is a function m on $\mathcal{C}(K) \times \mathbb{R}^+$ defined by

$$m(f, \delta) := \sup\{|f(x) - f(y)| : x, y \in K, d(x, y) < \delta\}$$

To say that f is uniformly continuous is to say that $\lim_{\delta \rightarrow 0} m(f, \delta) = 0$, and to say that \mathcal{F} is uniformly equicontinuous is to say that $\lim_{\delta \downarrow 0} \sup_{f \in \mathcal{F}} m(f, \delta) = 0$.

Note that, for fixed $\delta > 0$, the function $m(\cdot, \delta) : \mathcal{C}[0, 1] \rightarrow \mathbb{R} : f \mapsto m(f, \delta)$ is continuous: Indeed, using the inequality $|a| - |b| \leq |a - b| \leq |a| + |b|$, we see that

$$\begin{aligned} m(f_1, \delta) - m(f_2, \delta) &\leq \sup_{x, y \in K, d(x, y) < \delta} \left[|f_1(x) - f_1(y)| - |f_2(x) - f_2(y)| \right] \\ &\leq \sup_{x, y \in K, d(x, y) < \delta} \left| (f_1(x) - f_2(x)) - (f_1(y) - f_2(y)) \right| \\ &\leq 2\|f_1 - f_2\|_\infty \end{aligned}$$

from which the result follows. Combined with Dini's lemma, we thus see that if $n \rightarrow \infty$, then $m(\cdot, \frac{1}{n}) \rightarrow 0$ uniformly.

2.8.2 Compactness: Arzelà–Ascoli Theorem*

The Arzelà–Ascoli Theorem characterizes relative compactness in the space $\mathcal{C}[0, 1]$:

Theorem 2.8.16 (Arzelà–Ascoli)

Let $\mathcal{F} \subseteq \mathcal{C}[0, 1]$. Then \mathcal{F} is relatively compact if and only if

(i) $\sup_{f \in \mathcal{F}} |f(0)| < \infty$, and

(ii) $\lim_{\delta \downarrow 0} \sup_{f \in \mathcal{F}} m(f, \delta) = 0$

(if and only if \mathcal{F} is uniformly bounded and uniformly equicontinuous.)

Proof: First suppose that \mathcal{F} is relatively compact, i.e. totally bounded. Then it is obviously bounded in $\mathcal{C}[0, 1]$, and so

$$\sup_{f \in \mathcal{F}} |f(0)| \leq \sup_{f \in \mathcal{F}} \|f\|_\infty < \infty$$

which yields (i). To see (ii), recall that the modulus of continuity $m(f, \delta)$ is continuous in f . Any continuous function $f \in \mathcal{C}[0, 1]$ is uniformly continuous, i.e. has $m(f, \frac{1}{n}) \rightarrow 0$ as $n \rightarrow \infty$. Define $F_n : \mathcal{C}(\mathcal{F}) \rightarrow \mathbb{R} : f \mapsto m(f, \frac{1}{n})$. Then $F_n \rightarrow 0$ pointwise. By Dini's lemma, since \mathcal{F} is compact, we must have $F_n \rightarrow 0$ uniformly, i.e. $\sup_{f \in \mathcal{F}} |F_n(f)| \rightarrow 0$, from which (ii) follows.

For the converse, suppose that (i), (ii) hold for $\mathcal{F} \subseteq \mathcal{C}[0, 1]$. We must show that \mathcal{F} is relatively compact. First, we show that \mathcal{F} is uniformly bounded, i.e. that there is $M < \infty$ such that $\|f\|_\infty \leq M$ for all $f \in \mathcal{F}$. By (ii), choose n large enough that $\sup_{f \in \mathcal{F}} m(f, \frac{1}{n}) < \infty$. For $t \in [0, 1]$ we have

$$|f(t)| \leq |f(0)| + \sum_{k=1}^n |f(\frac{kt}{n}) - f(\frac{(k-1)t}{n})|$$

and thus

$$M := \sup_{0 \leq t \leq 1} \sup_{f \in \mathcal{F}} |f(t)| < \infty$$

It follows that \mathcal{F} is uniformly bounded, as asserted.

To show that \mathcal{F} is relatively compact, it suffices to show that it is totally bounded, as $\mathcal{C}[0, 1]$ is complete. Let $\varepsilon > 0$. Let Y be a finite subset of $[-M, M]$ with the property that any element of $[-M, M]$ lies within a distance of $< \varepsilon$ of some element of Y . Also choose n sufficiently large that $m(f, \frac{1}{n}) < \varepsilon$ for all $f \in \mathcal{F}$. Let \mathcal{P} be the set of all piecewise linear continuous functions, which are linear on each interval $I_{kn} := [\frac{k-1}{n}, \frac{k}{n}]$ with endpoints taking values in Y . The set \mathcal{P} is clearly finite, as Y is finite, and there are only finitely many intervals $I_{kn}, k = 1, \dots, n$. We claim that

$$\mathcal{F} \subseteq \bigcup_{p \in \mathcal{P}} B(p, 2\varepsilon)$$

Indeed, if $f \in \mathcal{F}$, then there is $p \in \mathcal{P}$ such that $|f(\frac{k}{n}) - p(\frac{k}{n})| < \varepsilon$ for $k = 0, \dots, n$. If $t \in [0, 1]$, there is k such that $t \in I_{kn}$. Then $p(t)$, being linear on I_{kn} , is a convex combination of the endpoints: $p(t) = \lambda p(\frac{k-1}{n}) + (1 - \lambda)p(\frac{k}{n})$ for some $0 \leq \lambda \leq 1$. Now

$$|f(t) - p(\frac{k}{n})| \leq |f(t) - f(\frac{k}{n})| + |f(\frac{k}{n}) - p(\frac{k}{n})| \leq 2\varepsilon$$

and similarly $|f(t) - p(\frac{k-1}{n})| < 2\varepsilon$ as well. Thus

$$|f(t) - p(t)| \leq \lambda |f(t) - p(\frac{k-1}{n})| + (1 - \lambda) |f(t) - p(\frac{k}{n})| < 2\varepsilon$$

This holds for all $t \in [0, 1]$, so $\|f - p\|_\infty < 2\varepsilon$ and hence $f \in B(p, 2\varepsilon)$, as required.

—

2.8.3 Separability: Stone–Weierstrass Theorem*

Theorem 2.8.17 (Weierstrass Approximation Theorem) *Suppose that $f : [a, b] \rightarrow \mathbb{R}$ is continuous, and let $\varepsilon > 0$. Then there exists a polynomial $B(x)$ such that $\|f - B\|_\infty < \varepsilon$*

Proof: We give a probabilistic proof, due to Bernstein. By scaling and translation, we may assume without loss of generality that the interval $[a, b]$ is the unit interval $[0, 1]$. For $p \in [0, 1]$, consider independent tosses of a coin which lands H with probability p . Let $X_k = 1$ if the k^{th} toss lands H, and let $X_k = 0$ otherwise. Let $S_n := \sum_{k \leq n} X_k$ be the number of H observed in n tosses. All of this is modeled on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then define

$$\begin{aligned} \mathbb{E}[f(\frac{S_n}{n})] &= \sum_{k=0}^n f(\frac{k}{n}) \mathbb{P}(S_n = k) \\ &= \sum_{k=0}^n f(\frac{k}{n}) \binom{n}{k} p^k (1-p)^{n-k} =: B_n(p) \end{aligned}$$

Now

$$|B_n(p) - f(p)| = |\mathbb{E}[f(\frac{S_n}{n}) - f(p)]|$$

and the law of large numbers implies $f(\frac{S_n}{n}) \rightarrow p$ a.s. This suggests a proof:

Use compactness of $[0, 1]$ and uniform continuity of f to choose $K, \delta > 0$ so that

$$\sup_x |f(x)| \leq K \quad \sup_{|x-y| < \delta} |f(x) - f(y)| < \frac{1}{2}\varepsilon$$

Define also

$$D_n := |f(\frac{S_n}{n}) - f(p)| \quad F_n := \{\omega \in \Omega : |\frac{S_n(\omega)}{n} - p| < \delta\}$$

Note that $\omega \in F_n$ implies $0 \leq D_n(\omega) \leq \frac{1}{2}\varepsilon$ and that $D_n(\omega) \leq 2K$ for all $\omega \in \Omega$. Further observe that Chebyshev's inequality yields

$$\mathbb{P}(F_n^c) \leq \frac{1}{\delta^2} \text{Var}(\frac{S_n}{n}) = \frac{1}{n^2 \delta^2} \text{Var}(\sum_{k \leq n} X_k) = \frac{1}{n \delta^2} (p - p^2) \leq \frac{1}{n \delta^2}$$

Thus

$$\begin{aligned} |B_n(p) - f(p)| &\leq \mathbb{E} D_n \\ &= \mathbb{E}[D_n; F_n] + \mathbb{E}[D_n; F_n^c] \\ &\leq \frac{1}{2}\varepsilon + 2K \mathbb{P}(F_n^c) \\ &= \frac{1}{2}\varepsilon + \frac{2K}{n \delta^2} \end{aligned}$$

Let $N \in \mathbb{N}$ be least so that $\frac{1}{2}\varepsilon + \frac{2K}{N \delta^2} < \varepsilon$. Thus $|B_n(p) - f(p)| < \varepsilon$ for all $p \in [0, 1]$ when $n \geq N$, and hence $B_n \rightarrow f$ uniformly in p .

—

Recall that if X is a compact topological space, then $\mathcal{C}(X)$ is the set of all continuous real-valued functions on X , equipped with the sup-norm $\|\cdot\|_\infty$. Further recall that

$$\max\{f, g\} = \frac{1}{2}(f+g+|f-g|) \quad \min\{f, g\} = -\max\{-f, -g\} \quad |f| = \max\{f, 0\} + \max\{-f, 0\}$$

Proposition 2.8.18 *Let K be a compact Hausdorff space, and let $\mathcal{F} \subseteq \mathcal{C}(K)$ be a closed set (w.r.t. the sup-norm). Suppose that*

- (1) \mathcal{F} contains all the constant functions;
- (2) \mathcal{F} is closed under addition and multiplication;
- (3) \mathcal{F} is point separating, i.e. if $x \neq y \in K$, then there is $f \in \mathcal{F}$ such that $f(x) \neq f(y)$;
- (4) \mathcal{F} is closed under absolute values, i.e. $f \in \mathcal{F}$ implies $|f| \in \mathcal{F}$.

Then $\mathcal{F} = \mathcal{C}(K)$.

Proof: Note that by (1), (2), (4) and the remarks preceding the statement of the theorem, that $f, g \in \mathcal{F}$ implies $\max\{f, g\}$ and $\min\{f, g\}$ belong to \mathcal{F} .

Fix $h \in \mathcal{C}(K)$, and $x \in K$. We first show that if $y \in K$, then there is $h_y \in \mathcal{F}$ such that

$$h_y(x) = h(x) \quad h_y(y) = h(y)$$

This is obvious if $x = y$. If $x \neq y$, choose any $f \in \mathcal{F}$ such that $f(x) \neq f(y)$ — this is possible because \mathcal{F} is point separating — and define

$$a := \frac{h(x) - h(y)}{f(x) - f(y)} \quad b := h(x) - af(x)$$

Observe that $h_y := af + b$ has the required properties, and that (1), (2) imply that $h_y \in \mathcal{F}$.

Given $\varepsilon > 0$, the sets $U_y := \{v \in K : h_y(v) > h(v) - \varepsilon\}$ form an open cover $\{U_y : y \in K\}$ of K , as $y \in U_y$. Since K is compact, there are $y_1, \dots, y_n \in K$ such that U_{y_1}, \dots, U_{y_n} cover K also. Now define

$$g_x := \max\{h_{y_k} : k = 1, \dots, n\}$$

Then $g_x \in \mathcal{F}$, with

$$g_x(x) = h(x) \quad g_x(v) > h(v) - \varepsilon \text{ for all } v \in K$$

This can be done for every $x \in K$. Now let $V_x := \{v \in K : g_x(v) < h(v) + \varepsilon\}$. Then $\{V_x : x \in K\}$ forms an open cover of K , as $x \in V_x$. By compactness, there is a finite subcover V_{x_1}, \dots, V_{x_m} . Define

$$g := \min\{g_{x_j} : j = 1, \dots, m\}$$

Then $g \in \mathcal{F}$. It follows easily that

$$h(v) - \varepsilon < g(v) < h(v) + \varepsilon \quad \text{for all } v \in K$$

and thus that $\|g - h\|_\infty < \varepsilon$.

Hence for every $h \in \mathcal{C}(K)$ and every $\varepsilon > 0$, we can find $g \in \mathcal{F}$ such that $\|g - h\|_\infty < \varepsilon$. Since \mathcal{F} is closed, it follows that $h \in \mathcal{F}$ for every $h \in \mathcal{C}(K)$.

Theorem 2.8.19 (Stone–Weierstrass Theorem)

Let K be a compact Hausdorff space, and let $\mathcal{F} \subseteq \mathcal{C}(K)$ be such that

- (1) \mathcal{F} contains all the constant functions;
- (2) \mathcal{F} is closed under addition and multiplication;
- (3) \mathcal{F} is point separating, i.e. if $x \neq y \in K$, then there is $f \in \mathcal{F}$ such that $f(x) \neq f(y)$.

Then \mathcal{F} is dense in $\mathcal{C}(K)$, i.e. every $h \in \mathcal{C}(K)$ is a uniform limit of members of \mathcal{F} .

Proof: Let $\bar{\mathcal{F}}$ be the closure of \mathcal{F} in $\mathcal{C}(K)$ (w.r.t. the sup-norm). Then $\bar{\mathcal{F}}$ satisfies (1), (2), (3) as well. By Propn. 2.8.18 it suffices to show that $\bar{\mathcal{F}}$ is closed under absolute values.

Note that (1), (2) imply that if P is a real polynomial and $f \in \bar{\mathcal{F}}$, then $P \circ f \in \bar{\mathcal{F}}$ (where $(P \circ f)(x) = P(f(x))$ for $x \in K$). Now if $f \in \bar{\mathcal{F}} \subseteq \mathcal{C}(K)$, then $M := \|f\|_\infty < \infty$, as K is compact. The function $A : [-M, M] \rightarrow \mathbb{R} : t \mapsto |t|$ is continuous, and is therefore a uniform limit of polynomials P_n , by Thm. 2.8.17. It follows that $|f| = A(f)$ is a uniform limit of $P_n \circ f \in \bar{\mathcal{F}}$. Thus $|f| \in \bar{\mathcal{F}}$ whenever $f \in \bar{\mathcal{F}}$. Thus $\bar{\mathcal{F}}$ satisfies (1)-(4) of Propn. 2.8.18.

—

Chapter 3

Motivation for Measure Theory

3.1 What is “Area”?

“Area” is a number associated with certain subsets of the Euclidean plane \mathbb{R}^2 , i.e. it is a function $|\cdot|$ which assigns to a set $E \subseteq \mathbb{R}^2$ its area $|E|$. Intuitively, the area function $|\cdot|$ should have certain properties:

1. If $E \subseteq \mathbb{R}^2$ is bounded, then $0 \leq |E| < \infty$.
2. $|\cdot|$ is *rotation- and translation invariant*: If a set F is obtained by rotating and shifting a set E , then $|F| = |E|$.
3. If E is a rectangle, then $|E| = \text{length} \times \text{breadth}$.
4. $|\cdot|$ is *additive*: If E, F are disjoint bounded subsets of \mathbb{R}^2 , then $|E \cup F| = |E| + |F|$.
More generally, if E_1, E_2, \dots is a sequence of mutually disjoint bounded subsets \mathbb{R}^2 , then $|\bigcup_n E_n| = \sum_{n=1}^{\infty} |E_n|$.

It follows easily how to calculate the area of triangles, and thus that of polygons. But what about non-polygonal sets? For example, how do we justify that the area of a circle of radius r is πr^2 ? Before we do this, do the following exercise.

Exercise 3.1.1 Use the properties (1)-(4) of the area function to show the following:

- (a) $|\emptyset| = 0$.
- (b) If $A \subseteq B$ then $|B - A| = |B| - |A|$.
- (c) If $A \subseteq B$ then $|A| \leq |B|$.
- (d) If $A_1 \subseteq A_2 \subseteq A_3 \subseteq \dots$ and if $A := \bigcup_n A_n$, then $|A| = \lim_n |A_n|$.
[Hint: Define $B_1 := \emptyset$, and for $n = 1, 2, 3, \dots$, let $B_{n+1} := A_{n+1} - A_n$. Then $A_n = \bigcup_{j=1}^n B_j$ is a union of disjoint sets. Use the fact that the value infinite series is a limit of finite sums.]

□

Exercise 3.1.2 We use the properties of the area function to show that the area of an open circle A of radius r is $|A| = \pi r^2$.

- (a) For $n = 1, 2, 3, \dots$, let A_n be the regular open polygon with 2^{n+1} sides, inscribed in a circle of radius r . Thus A_1 is a square, A_3 an octagon, etc. Note that $A_1 \subseteq A_2 \subseteq A_3 \subseteq \dots$. Also note that $\bigcup_n A_n = A$. (This is why we need the sets A_n and A to be open subsets of \mathbb{R}^2).

- (b) A_n consists of 2^{n+1} congruent isosceles triangles, constructed by joining each of the sides of the polygon to the centre of the circle. Explain why each such triangle has area $\frac{1}{2}r^2 \sin \frac{\pi}{2^n}$, and conclude that $|A_n| = 2^n r^2 \sin \frac{\pi}{2^n}$.
- (c) Conclude that $|A| = \lim_n \pi r^2 \left(\frac{\sin \frac{\pi}{2^n}}{\frac{\pi}{2^n}} \right) = \pi r^2$.

□

The technique used to compute the area in Exercise 3.1.2 relies on the set A being approximated *from the inside* by triangles. Not every set can be so approximated, however. Take for example the set $A := \{(p, q) : p, q \text{ are rational numbers with } 0 \leq p, q \leq 1\}$. It is not clear what the area of A should be: On the one hand, the set A is dense inside the unit rectangle, so one might guess that $|A| = 1$. If we approximate A *from the outside* however, we obtain a convincing argument that $|A| = 0$:

Exercise 3.1.3 Recall that the set of rational numbers is countable. So we can write the elements of A in a list: $A = \{(p_n, q_n) : n \in \mathbb{N}\}$. Fix $\varepsilon > 0$. For $n \in \mathbb{N}$, let R_n be a square centred at (p_n, q_n) with sides $\frac{\varepsilon}{2^{n/2}}$. Let $B = \bigcup_n R_n$, and show that $|B| \leq \sum_n |R_n| = \varepsilon$. Also show that $A \subseteq B$ so that $|A| \leq \varepsilon$. Since $\varepsilon > 0$ was arbitrary, we have $|A| \leq \varepsilon$ for all $\varepsilon > 0$, i.e. $|A| = 0$.

□

Observe that the technique used in Exercise 3.1.3 can be used to prove that *every* countable subset of \mathbb{R}^2 has zero area. It is therefore necessary that the set of real numbers be uncountable for the concept of area to have a useful meaning!

Later in this course we shall show that it is impossible to assign an area to *every* bounded subset of \mathbb{R}^2 , i.e. there is no function which satisfies each of the properties (1)-(4) of area above, and which is defined for every bounded subset of \mathbb{R}^2 . Thus there are subsets of \mathbb{R}^2 which have no area. This does *not* mean that these sets have zero area; it means that there is no number which can be called their area, and which is consistent with (1)-(4).

Remarks 3.1.4 Just so, some subsets of \mathbb{R}^3 fail to have a volume. This is illustrated by the following theorem, called the *Banach–Tarski paradox*: It is possible to break up a solid ball the size of a pea into finitely many pieces, and to rearrange these pieces to form a solid ball the size of the sun (or into pretty much any shape of any size that you desire)¹. [However, the proof of this theorem depends on an axiom about sets that most people find obvious, but that mathematicians of the early twentieth century deemed highly controversial, namely the *axiom of choice*. This axiom states that if you have a family of non-empty disjoint sets A_i ($i \in I$), then there is a set B which contains exactly one element from each A_i .]

□

3.2 Shortcomings of the Riemann Integral

We recall briefly how the Riemann integral $\int_a^b f(t) dt$ is defined: Let f be a real-valued function defined and bounded on an interval $[a, b]$. A *partition* P of $[a, b]$ is a finite ordered set $\{a = t_0 < t_1 < t_2 < \cdots < t_n = b\}$. The *size* of such a partition is denote $\sigma(P)$, and defined by

$$\sigma(P) := \max_k (t_k - t_{k-1})$$

¹No, really, this is a **theorem**.

A *tagged partition* is a partition P together with a choice $t_k^* \in [t_{k-1}, t_k]$ for each $k = 1, \dots, n$. Tagged partitions will be indicated by a $*$, i.e. if P is a partition, then P^* denotes an associated tagged partition.

With each tagged partition, we can associate a Riemann sum

$$S(P^*, f) := \sum_{k=1}^n f(t_k^*) (t_k - t_{k-1}) = \sum_{k=1}^n f(t_k^*) \Delta_k t$$

The Riemann integral $\int_a^b f \, dt$ should be the limit of the Riemann sums, over all tagged partitions P^* , as $\sigma(P) \rightarrow 0$. To be precise, we say

$$\lim_{\sigma(P) \rightarrow 0} S(P^*, f) = L \text{ exists}$$

if and only if for every $\varepsilon > 0$ there is $\delta > 0$ such that

$$|S(P^*, f) - L| < \varepsilon \quad \text{whenever} \quad \sigma(P) < \delta$$

Then we define

$$\int_a^b f \, dt := \lim_{\sigma(P) \rightarrow 0} S(P^*, f)$$

provided this limit exists, and say that f is Riemann integrable with respect to G on $[a, b]$.

With each partition $\{a = t_0 < t_1 < \dots < t_n = b\}$ it is possible to associate three natural tagged partitions, namely those having tags equal to the left endpoint, right endpoint and midpoint of each interval. This yields:

- The lefthand Riemann sum $\sum_k f(t_{k-1}) \Delta_k t$;
- The righthand Riemann sum $\sum_k f(t_k) \Delta_k t$;
- The symmetric Riemann sum $\sum_k f(\frac{t_{k-1} + t_k}{2}) \Delta_k t$.

If f is Riemann integrable over $[a, b]$, then each of these sums must converge as $\sigma(P) \rightarrow 0$, and all to the same limit.

Remarks 3.2.1 A slightly different definition uses *Darboux sums* rather than Riemann sums. Given a real-valued functions f defined and bounded on an interval $[a, b]$, and a partition $P = \{a = t_0 < t_1 < \dots < t_n = b\}$, let the upper and lower Darboux sums be defined by

$$U(P, f) := \sum_{k=1}^n \sup\{f(t) : t \in [t_{k-1}, t_k]\} \cdot (t_k - t_{k-1})$$

$$L(P, f) := \sum_{k=1}^n \inf\{f(t) : t \in [t_{k-1}, t_k]\} \cdot (t_k - t_{k-1})$$

If f is continuous on $[a, b]$, it attains its supremum and infimum on each subinterval, i.e. we can choose $t_k^{\max}, t_k^{\min} \in [t_{k-1}, t_k]$ such that

$$f(t_k^{\max}) = \sup\{f(t) : t \in [t_{k-1}, t_k]\} \quad f(t_k^{\min}) = \inf\{f(t) : t \in [t_{k-1}, t_k]\}$$

If $P^{*\max}, P^{*\min}$ are the tagged partitions given by $a = t_0 < \dots < t_n = b$ and the tags t_k^{\max}, t_k^{\min} respectively, then it is easy to see that

$$U(P, f) = S(P^{*\max}, f) \quad L(P, f) = S(P^{*\min}, f)$$

i.e. the Darboux sums give the most extreme values of the Riemann sums for any given partition. However, the Darboux sums may differ from Riemann sums if f is not continuous.

The Riemann sums may be defined even when f is Banach space-valued, however, whereas the Darboux sums, being dependent on sup's and inf's, make sense for real-valued functions only.

□

From an earlier course in analysis, we know that the Riemann integral $\int_a^b f \, dt$ exists when f is continuous (or even piecewise continuous) on $[a, b]$. When the function is too discontinuous, we run into trouble, however:

Example 3.2.2 Consider the *Dirichlet function*

$$I_{\mathbb{Q}}(t) := \begin{cases} 1 & \text{if } t \in \mathbb{Q} \\ 0 & \text{else} \end{cases}$$

where \mathbb{Q} is the set of rational numbers. If $P = \{a = t_0 < t_1 < \cdots < t_n = b\}$ is any partition of $[a, b]$, no matter how fine, we can always find tags $t_k^*, t'_k \in [t_{k-1}, t_k]$ so that t_k^* is rational, and t'_k is irrational. Thus $I_{\mathbb{Q}}(t_k^*) = 1, I_{\mathbb{Q}}(t'_k) = 0$. It follows that

$$S(P^*, f) = \sum_k 1 \cdot (t_k - t_{k-1}) = 1 - 0 = 1 \quad S(P', f) = \sum_k 0 \cdot (t_k - t_{k-1}) = 0$$

and thus $S(P^*, f), S(P', f)$ cannot be made to lie arbitrarily close to each other, no matter how fine the partition P . Thus $\lim_{\sigma(P) \rightarrow 0} S(P, f)$ does not exist.

□

When the Riemann integral is first encountered, it is taught as “the area under a curve”: If $f \geq 0$ is continuous, then $\int_a^b f \, dt$ is the area under the curve described by f , between $t = a$ and $t = b$. For $A \subseteq \mathbb{R}$, define the *indicator function* of A by

$$I_A(t) := \begin{cases} 1 & \text{if } t \in A \\ 0 & \text{else} \end{cases}$$

Consider now the $I_{\mathbb{Q}}$, where \mathbb{Q} is the set of rational numbers. This function is *very discontinuous*. If we try to compute this “curve” over the interval $[0, 1]$ using the Riemann integral, we run into trouble: The Riemann integral $\int_0^1 I_{\mathbb{Q}} \, dt$ does not exist.

We can make a convincing argument that the area under the curve over the interval $[0, 1]$ should be zero, as follows: Use the fact that \mathbb{Q} is countable to enumerate the rational numbers in $[0, 1]$, i.e. write $[0, 1] \cap \mathbb{Q} = \{q_n : n \in \mathbb{N}\}$. For any $\varepsilon > 0$, define

$$B_n := [q_n - \frac{\varepsilon}{2^{n+1}}, q_n + \frac{\varepsilon}{2^{n+1}}] \text{ for } n \in \mathbb{N}, \quad f = I_{\bigcup_n B_n}$$

The area under the curve of f is made up of (possibly overlapping) rectangles of height 1 centered at the rational numbers. Thus the area under f over $[0, 1]$ is $\leq \sum_n 1 \cdot (\text{length of } B_n) = \sum_n \frac{\varepsilon}{2^n} = \varepsilon$. It is also clear that $0 \leq I_{\mathbb{Q}} \leq f$, and thus that the area under $I_{\mathbb{Q}}$ is less than the area under f , i.e. that the area under $I_{\mathbb{Q}}$ is $\leq \varepsilon$. Since this is true for any $\varepsilon > 0$, we conclude that the area under $I_{\mathbb{Q}}$ is 0.

Thus we have the following:

$$\int_0^1 I_{\mathbb{Q}} \, dt \text{ is } \textit{undefined}, \text{ but it } \textit{should} \text{ be zero}$$

The Riemann integral is simply not powerful enough to handle functions like $I_{\mathbb{Q}}$.

You may counter that a function such as $I_{\mathbb{Q}}$ is *pathological*, and unlikely to be encountered in practice. It is true that we chose it here simply to make a point. However, the following example should cause you to feel uneasy about the assertion that $I_{\mathbb{Q}}$ is “pathological”:

Example 3.2.3 Consider the function $g(t)$ defined by

$$g(t) = \lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \cos(m! \pi t)^{2n}$$

If t is a rational number, i.e. $t = \frac{p}{q}$ where $p \in \mathbb{Z}, q \in \mathbb{N}$, then $m! t \in \mathbb{Z}$ for all $m \geq q$. Consequently, $\cos(m! \pi t) = \pm 1$ for all $m \geq q$, and thus $\cos(m! \pi t)^{2n} = 1$ for all n and all $m \geq q$. It follows that $g(t) = 1$ when t is rational.

On the other hand, if t is irrational, then $0 \leq |\cos(m! \pi t)| < 1$ for all m , and so $0 \leq \cos(m! \pi t)^{2n} < 1$ for all m . Now if $0 < x < 1$, then $x^n \rightarrow 0$. It follows that $\lim_{n \rightarrow \infty} \cos(m! \pi t)^{2n} = 0$ for all m , and thus that $g(t) = 0$ when t is irrational. Hence

$$I_{\mathbb{Q}} = \lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \cos(m! \pi t)^{2n}$$

The “pathological” function $I_{\mathbb{Q}}$ therefore appears as a limit of perfectly ordinary functions.

□

This brings us to the next problem: The Riemann integral does not handle limits well. There are many other examples of functions $f_n \rightarrow f$, where f_n are perfectly good Riemann integrable functions, and where $\int_a^b f \, dt$ *should* be $\lim_n \int_a^b f_n \, dt$, but where f fails to be Riemann integrable. A *better* integral is called for.

3.3 Motivation from Probability Theory

A model for an experiment involving randomness takes the form $(\Omega, \mathcal{F}, \mathbb{P})$. Intuitively, Ω is the set of all possible outcomes of the experiment, and is called the *sample space*. \mathcal{F} is the set of all *events*, i.e. “permissible” combinations of outcomes. (We shall see that not all combinations of events are permissible – it is simply impossible to create a consistent mathematical theory of probability in which every set of outcomes is permissible.) \mathbb{P} is a map $\mathcal{F} \xrightarrow{\mathbb{P}} [0, 1]$ which assigns to each (permissible) event its probability.

Example 3.3.1 A die is rolled once. The possible outcomes are an integer between one and six. Thus the sample space can be taken to be $\Omega = \{1, 2, \dots, 6\}$. We may be interested in the following events:

- (a) The outcome is the number 1;
- (b) The outcome is an even number;
- (c) The outcome is an odd number which is strictly greater than 1;

Each of these events can be described by a subset of the sample space. Thus if A, B, C are the subsets corresponding to the events (a), (b), (c), then

$$\begin{aligned} A &= \{1\} \\ B &= \{2, 4, 6\} \\ C &= \{3, 5\} \end{aligned}$$

The probabilities of these events, by elementary reasoning, are $\mathbb{P}(A) = \frac{1}{6}$, $\mathbb{P}(B) = \frac{1}{2}$ and $\mathbb{P}(C) = \frac{1}{3}$, provided that the die is fair. Every subset of Ω is a “permissible” event, and thus $\mathcal{F} = \mathcal{P}(\Omega)$.

□

Mathematically, an *event* is a *set*, i.e. events are just subsets of the sample space. The outcome of any random experiment must be some element ω of the sample space Ω . Now Ω is itself a subset of Ω and thus corresponds to some event. We call it the *certain event*, since we are certain that $\omega \in \Omega$. We must always have $\mathbb{P}(\Omega) = 1$. The empty set \emptyset is also a subset of Ω and thus corresponds to some event. We call it the *impossible event*, since it is impossible that an outcome ω is in \emptyset . We will always have $\mathbb{P}(\emptyset) = 0$.

Note that the sample space corresponding to a random experiment need not be unique. Consider, for example, the random experiment of rolling two dice. Then we can choose the sample space to be the 36 element set $\Omega_1 = \{(i, j) : i, j \text{ positive integers between 1 and 6}\}$. The probabilities for each outcome are then the same: $\mathbb{P}(\omega) = \frac{1}{36}$ for each $\omega \in \Omega_1$. This is a so-called *uniform distribution*. On the other hand, we can choose the sample space to be the 11-element set $\Omega_2 = \{2, 3, 4, \dots, 12\}$ corresponding to the total of the two dice. In this case, the probability distribution is non-uniform: $\mathbb{P}(\{7\}) = \frac{1}{6}$ whereas $\mathbb{P}(\{2\}) = \frac{1}{36}$. Choosing the sample space and the corresponding probability distribution for a particular situation is part of the art of probabilistic modelling.

Example 3.3.2 A coin is flipped until the first head turns up. This may happen on the first toss or the second, or...or never. Thus the sample space is $\Omega = \{\omega_1, \omega_2, \dots, \omega_\infty\}$, where the outcome ω_n denotes the event that the first head turns up on the n^{th} toss, and ω_∞ denotes the event of never flipping heads. It is clear from elementary probability that $\mathbb{P}(\{\omega_n\}) = \frac{1}{2^n}$ (provided that the coin is fair). We may now consider various composite events, such as:

- (a) Let A be the event that the first head appears on either the third or the fourth toss. Then $A = \{\omega_3\} \cup \{\omega_4\} = \{\omega_3, \omega_4\}$. Clearly $\mathbb{P}(A) = \frac{1}{2^3} + \frac{1}{2^4}$.
- (b) Let B be the event that the first head appears after an even number of tosses. Thus $B = \bigcup_{n \in \mathbb{N}} \{\omega_{2n}\}$ and $\mathbb{P}(B) = \sum_{n=1}^{\infty} \frac{1}{2^{2n}} = \frac{1}{3}$. Did you think that the probability that the first head appears after an even number of tosses is $\frac{1}{2}$? If so, note that the probability that the first head appears on the first toss is $\frac{1}{2}$, and the probability that the first head appears after an odd number of tosses is therefore greater than $\frac{1}{2}$.
- (c) Let C be the event that *both* events A and B occur. Clearly $C = \{\omega_4\} = A \cap B$.
- (d) Let D be the event that a head does occur after a finite number of tosses. Thus D is the complement of the event that heads never occurs. Thus $D = \Omega - \{\omega_\infty\} = \{\omega_1, \omega_2, \dots\}$. Hence $\mathbb{P}(D) = \sum_{n=1}^{\infty} \frac{1}{2^n} = 1$. This can also be seen from the fact that $\mathbb{P}(\{\omega_\infty\}) = 0$.

□

3.4 Structure of Events

In order for our mathematical theory of probability to bear some resemblance to the real world, it is clear that we should be able to combine events in the following ways:

- If A is an event, then the possibility of A not occurring should also be an event. Now if the outcome of a random experiment is $\omega \in \Omega$, then the event A occurs if and only if $\omega \in A$ (remember that we consider an event to be a subset of the sample space). Thus the event that A does not occur corresponds to $\omega \notin A$, i.e. to the set $A^c = \Omega - A$. We want the probabilities of these events to be related by $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
- If A, B are events, then the possibility of both A and B occurring should also be an event. Now if the outcome of a random experiment is $\omega \in \Omega$, then both A and B occur if and only if $\omega \in A$ and $\omega \in B$, i.e. if and only if $\omega \in A \cap B$. Thus the event of both A and B occurring corresponds to the set $A \cap B$.
- In the same way, if A and B are events, then the possibility of at least one of A or B occurring should be an event as well. This corresponds to the set $A \cup B$. We say that events are *disjoint* or *mutually exclusive* if they cannot occur simultaneously. Thus if A, B are disjoint, the $\omega \in A$ implies $\omega \notin B$. Clearly, therefore, A and B are disjoint

if and only if $A \cap B = \emptyset$ (i.e. the event that both A and B occur is impossible). For disjoint events A and B , we want $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$. This is because $\mathbb{P}(A \cup B) = \frac{N_{A \cup B}}{N} = \frac{N_A + N_B}{N} = \mathbb{P}(A) + \mathbb{P}(B)$ (where N_A is the number of elements in the set A , etc.)

- In fact we demand more: Given a countable list of events A_1, A_2, A_3, \dots , the possibilities of either all of these events occurring, or of at least one of these events occurring, should also be events. They correspond to the sets $\bigcap_{n=1}^{\infty} A_n$ and $\bigcup_{n=1}^{\infty} A_n$ respectively. If the events A_n are mutually disjoint, i.e. if $A_n \cap A_m = \emptyset$ whenever $n \neq m$, then we want
$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$$

The concept of *probability* has rather a lot in common with that of *area*:

- “Probability” is measured by non-negative number $\mathbb{P}(A)$ assigned to a subset A of the sample space Ω .
“Area” is measured by non-negative number $|A|$ assigned to a subset of \mathbb{R}^2 .
- $\mathbb{P}(\emptyset) = 0$;
 $|\emptyset| = 0$.
- If A_n are disjoint events, then $\mathbb{P}(\bigcup_n A_n) = \sum_n \mathbb{P}(A_n)$;
If A_n are disjoint sets, then $|\bigcup_n A_n| = \sum_n |A_n|$.

When we isolate and study the common features of probability and area, we get the subject of *measure theory*. We shall show that we can develop a theory which allows us to form integrals $\int f d\mu$ of functions f with respect to *measures* μ (rather than variables). It will turn out that the integral with respect to *Lebesgue measure* (yet to be defined) is precisely the more powerful generalization of the Riemann integral that we seek. It will also transpire that the integral with respect to a probability measure precisely captures the notion of *probabilistic expectation*.

Armed with the intuition and motivation provided by the above examples, we now proceed with the formal theory.

Chapter 4

Measure Spaces

4.1 Events and σ -algebras

To model a *random experiment*, we need to define three objects:

- A *sample space* Ω , representing the possible *outcomes* of the experiment. The outcomes $\omega \in \Omega$ are called *sample points*.
- A family \mathcal{F} of *events*.
An event is a (permissible/relevant) subset of Ω . If A is an event, we say that A occurs if the outcome ω is an element of A .
We shall require \mathcal{F} to be a σ -*algebra* (which we define below).
- A *probability measure* \mathbb{P} which assigns to each event a probability.
Thus $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$.

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ will be called a *probability space* (subject to certain conditions on \mathcal{F} and \mathbb{P}).

Recall that an event $E \in \mathcal{F}$ is said to have occurred if the outcome ω of the random experiment belongs to E . Intuitively, we think of \mathcal{F} as a set of events E for which we can decide whether or not E occurred at the termination of the experiment. Note: *whether* or *not*.

This intuition imposes the following constraints on \mathcal{F} :

- (a) $\Omega \in \mathcal{F}$ and $\emptyset \in \mathcal{F}$.
Indeed, every outcome ω belongs to Ω , and thus the event Ω always occurs — it's the *certain* event.
Similarly, no outcome ω belongs to \emptyset , and thus the event \emptyset never occurs — it's the *impossible* event.
- (b) If $E \in \mathcal{F}$, then $E^c \in \mathcal{F}$, i.e. \mathcal{F} is *closed under complementation*.
For if we can decide whether or not E occurred, then we can also decide whether or not E^c occurred: For suppose that the outcome of the experiment is ω . If E occurred, then $\omega \in E$, so $\omega \notin E^c$, hence E^c did not occur.
Similarly, if E did not occur, then E^c did occur.

- (c) If $E_1, E_2 \in \mathcal{F}$, then $E_1 \cap E_2 \in \mathcal{F}$, i.e. \mathcal{F} is *closed under intersection*.
 For if we can decide whether or not E_1 occurred, and also whether or not E_2 occurred, then we can decide whether or not $E_1 \cap E_2$ occurred: $E_1 \cap E_2$ occurred iff $\omega \in E_1 \cap E_2$ iff $\omega \in E_1$ and $\omega \in E_2$ iff *both* E_1 and E_2 occurred.
 Thus if we can decide whether or not E_1, E_2 occurred, we can also decide whether or not $E_1 \cap E_2$ occurred.
- (d) Similarly, \mathcal{F} is *closed under union*: The event $E_1 \cup E_2$ occurs iff either E_1 occurred, or E_2 occurred (or both).
- (e) We can generalize (c) and (d) somewhat: If $E_1, E_2, E_3, \dots, E_n, \dots$, is a countable sequence of members of \mathcal{F} , then also $\bigcap_n E_n \in \mathcal{F}$ and $\bigcup_n E_n \in \mathcal{F}$, i.e. \mathcal{F} is *closed under countable intersections and -unions*.
 For $\bigcap_n E_n$ occurred iff each of the E_n occurred, and $\bigcup_n E_n$ occurred iff at least one of the E_n occurred. Thus if we can decide whether or not each E_n occurred, we can also decide whether or not $\bigcap_n E_n$ and $\bigcup_n E_n$ occurred.

This leads to the following definitions:

Definition 4.1.1 Let Ω be a set. A collection \mathcal{A} of subsets of Ω is called an *algebra* (or *field*) on Ω if

- (i) $\emptyset \in \mathcal{A}$;
- (ii) $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$;
- (iii) $A, B \in \mathcal{A} \Rightarrow A \cup B \in \mathcal{A}$.

A collection \mathcal{F} of subsets of Ω is called a σ -*algebra* (or σ -*field*) if it satisfies (i), (ii) and

- (iii) $_{\sigma}$ If $A_n \in \mathcal{F}$ (for $n \in \mathbb{N}$), then $\bigcup_n A_n \in \mathcal{F}$.

Exercise 4.1.2 (i) An algebra is closed under (finite) intersections.

(ii) A σ -algebra is closed under countable intersections.

(iii) A σ -algebra is also an algebra.

(iv) If Ω is a finite set, then any algebra on Ω is also a σ -algebra.

(v) If \mathcal{A} is an algebra and $A, B \in \mathcal{A}$, then $A - B$ and $A \Delta B$ belong to \mathcal{A} .

item If Ω is a set, then $\mathcal{F}_0 = \{\emptyset, \Omega\}$ is the smallest σ -algebra on Ω , and $\mathcal{F}_{\infty} = \mathcal{P}(\Omega)$ is the biggest σ -algebra on Ω .

(vi) If $\{\mathcal{F}_i : i \in I\}$ is a family of σ -algebras on Ω , then $\mathcal{F} := \bigcap_{i \in I} \mathcal{F}_i$ is also a σ -algebra on Ω .

□

Events are organized in σ -algebras. The set-theoretic operations $\bigcup, \bigcap, \cdot^c$, correspond to logical combinations *or, and, not* of events.

Frequently, the events of interest form a collection \mathcal{C} which is not a σ -algebra. Suppose that \mathcal{C} is a collection of events which can be *decided*, i.e. if $E \in \mathcal{C}$, then we can decide whether or not E occurred. We can then also decide whether or not E^c occurred, but E^c may not be an element of \mathcal{C} . The bigger set \mathcal{F} of all events that can be decided, given that we can decide all the events in \mathcal{C} , is a σ -algebra containing \mathcal{C} .

Definition and Proposition 4.1.3 Let \mathcal{C} be a family of subsets of Ω . There exists a unique smallest σ -algebra \mathcal{F} which contains \mathcal{C} (i.e. $\mathcal{C} \subseteq \mathcal{F}$, and if \mathcal{G} is any σ -algebra such that $\mathcal{C} \subseteq \mathcal{G}$, then $\mathcal{F} \subseteq \mathcal{G}$ also).

\mathcal{F} is called the σ -algebra generated by \mathcal{C} , and denoted by

$$\mathcal{F} = \sigma(\mathcal{C})$$

Proof: Let $\mathbb{F} = \{\mathcal{G} : \mathcal{G} \text{ a } \sigma\text{-algebra with } \mathcal{C} \subseteq \mathcal{G}\}$, and let $\mathcal{F} = \bigcap \mathbb{F}$. Then $\mathcal{F} \in \mathbb{F}$. (Why?) Moreover, if \mathcal{G} is a σ -algebra which contains \mathcal{C} , then $\mathcal{G} \in \mathbb{F}$, and so $\mathcal{F} \subseteq \mathcal{G}$. (Why?)

□

We repeat the following important intuition

$\sigma(\mathcal{C})$ consists of all those events F for which we can *decide* whether or not F has occurred, given that we *know* exactly which of the $E \in \mathcal{C}$ have occurred.

Definition 4.1.4 If the sample space Ω is a topological space, we define the *Borel algebra* of Ω by

$$\mathcal{B}(\Omega) = \sigma(\text{open sets of } \Omega)$$

In particular, $\mathcal{B}(\mathbb{R})$ is the smallest σ -algebra on \mathbb{R} which contains all the open subsets of \mathbb{R} .

$\mathcal{B}(\mathbb{R})$ is one of the most important σ -algebras.

Exercise 4.1.5 Prove that the following sets belong to $\mathcal{B}(\mathbb{R})$:

- (i) All closed intervals $[x, y]$, where $x \leq y \in \mathbb{R}$.
- (ii) The half-open intervals $(x, y]$ and $[x, y)$, where $x < y \in \mathbb{R}$.
- (iii) Every singleton $\{x\}$, where $x \in \mathbb{R}$.
- (iv) Every countable subset of \mathbb{R} .
- (v) The sets \mathbb{Q} of rational numbers and Irr of irrational numbers.

□

Exercise 4.1.6 Define

$$\mathcal{C} := \text{collection of all intervals of the form } (-\infty, x], \quad \text{where } x \in \mathbb{R}$$

Show that $\sigma(\mathcal{C}) = \mathcal{B}(\mathbb{R})$.

(You may use the fact that every open subset of \mathbb{R} can be represented as a union of countably many open intervals.)

□

4.2 Measures

The notion of *measure* generalizes the concepts of length, area, volume, mass, and probability.

Definition 4.2.1 Let \mathcal{F} be a σ -algebra on a set ω . A function $\mu : \mathcal{F} \rightarrow \bar{\mathbb{R}}$ is called a (countably additive, non-negative) *measure* if and only if

- (i) μ is *non-negative*: $0 \leq \mu A \leq \infty$ for each $A \in \mathcal{F}$.
- (ii) $\mu \emptyset = 0$.
- (iii) μ is *countably additive* (or σ -additive): If $A_1, A_2, \dots \in \mathcal{F}$ is a countable sequence of *pairwise disjoint* sets, then

$$\mu\left(\bigcup_n A_n\right) = \sum_n \mu A_n$$

If $\mu\Omega = 1$, then μ is called a *probability measure*.

If \mathcal{F} is a σ -algebra on a set Ω , then the pair (Ω, \mathcal{F}) is called a *measurable space*. The elements of \mathcal{F} are called *measurable sets*, or *events* in the probabilistic framework. If, in addition, μ is a measure on \mathcal{F} , the triple $(\Omega, \mathcal{F}, \mu)$ is called a *measure space*. The symbols \mathbb{P}, \mathbb{Q} are used for probability measures, and $(\Omega, \mathcal{F}, \mathbb{P})$ will always denote a *probability space*.

Example 4.2.2 Important: Lebesgue Measure: We shall prove later there is a unique measure λ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, which assigns every interval its length, i.e.

$$\lambda(a, b) = \lambda(a, b] = \lambda[a, b] = b - a$$

This measure is called *Lebesgue measure*, which provided them original impetus for the development of the subject of measure theory.

Lebesgue measure is also important in probability theory. Consider, for example, the experiment of drawing a uniformly distributed random number from the unit interval $[0, 1]$. The probability of drawing a number $\geq \frac{1}{2}$ is $\mathbb{P}([\frac{1}{2}, 1]) = \frac{1}{2}$. The probability of drawing a number between $\frac{1}{4}$ and $\frac{1}{3}$ is $\mathbb{P}([\frac{1}{4}, \frac{1}{3}]) = \frac{1}{12}$. Similarly, the probability of drawing a number between a and b (where $0 \leq a \leq b \leq 1$) is $\mathbb{P}([a, b]) = b - a$. Thus the appropriate measure \mathbb{P} is just Lebesgue measure, restricted to $[0, 1]$.

There are higher dimensional analogues of Lebesgue measure: There is a measure, also denoted λ and called Lebesgue measure, on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, which assigns to every n -dimensional rectangle its volume.

□

Example 4.2.3 Suppose that $F : \mathbb{R} \rightarrow \mathbb{R}$ is an increasing right-continuous function, i.e.

$$F(s) \leq F(t) \quad \text{when } s \leq t \quad \text{and} \quad F(t) = \lim_{s \downarrow t} F(s)$$

We shall prove later that there is a unique measure μ_F on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with the property that

$$\mu_F(a, b] = F(b) - F(a) \quad \text{for all } a < b \in \mathbb{R}$$

μ_F is called the *Lebesgue-Stieltjes* measure associated with F . Note that if $F(t) := t$, then $\mu_F = \lambda$ is Lebesgue measure.

□

Exercise 4.2.4 Let (X, \mathcal{F}) be a measurable space, and let $x_0 \in X$. Define $\delta_{x_0} : \mathcal{F} \rightarrow \mathbb{R}$ by

$$\delta_{x_0}(F) = \begin{cases} 1 & \text{if } x_0 \in F \\ 0 & \text{if } x_0 \notin F \end{cases}$$

Show that δ_{x_0} is a measure on (X, \mathcal{F}) .

δ_{x_0} is called the *Dirac measure*, or *point mass*, at x_0 .

□

Note that, for general measures, we allow $+\infty$ as a value. For example, the length of the real line is $+\infty$, so $\lambda(\mathbb{R}) = +\infty$, where λ is Lebesgue measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. However, we often need to get a “handle” on infinity:

Definition 4.2.5 A measure μ on a measurable space (Ω, \mathcal{F}) is called

- (a) *finite*, if $\mu\Omega < \infty$;
- (b) *σ -finite*, if Ω is the countable union of sets of finite measure, i.e. if there is a sequence A_1, A_2, \dots of measurable sets such that each $\mu A_n < \infty$, and such that $\Omega = \bigcup_n A_n$.

Thus Lebesgue measure is σ -finite, but not finite (Why?).

Proposition 4.2.6 (Additivity Properties)

Suppose that $(\Omega, \mathcal{F}, \mu)$ is a measure space, and let $A, B, A_1, A_2, \dots \in \mathcal{F}$.

- (a) If $A \subseteq B$, then $\mu A \leq \mu B$.
- (b) If $A \subseteq B$ and $\mu A < \infty$, then $\mu(B - A) = \mu B - \mu A$.
- (c) $\mu(A \cup B) = \mu A + \mu B - \mu(A \cap B)$
- (d) $\mu(\bigcup_n A_n) \leq \sum_n \mu A_n$.

Exercise 4.2.7 (a) Prove the preceding proposition.

(b) Suppose that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, and that $A, A_1, A_2, \dots \in \mathcal{F}$.

- (i) Show that $\mathbb{P}(A^c) = 1 - \mathbb{P}A$.
- (ii) Show that if $\mathbb{P}A_n = 1$ for $n \in \mathbb{N}$, then $\mathbb{P}(\bigcap_n A_n) = 1$ also.

□

Next, we introduce some useful terminology:

Definition 4.2.8 Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, and let $A \subseteq \Omega$.

- (a) We say that A is μ -null if there exists $B \in \mathcal{F}$ such that $A \subseteq B$ and $\mu B = 0$.
- (b) We shall say that a statement φ holds μ -almost everywhere (or μ -almost surely in the probabilistic framework), if the set of $\omega \in \Omega$ where φ fails to hold is μ -null.

We abbreviate μ -almost everywhere and μ -almost surely by μ -a.e. and μ -a.s. respectively.

Remarks 4.2.9 Note that in (a) above, the set A might not belong to \mathcal{F} so μA might be undefined. However, μA “ought” to be zero. Later, this insight will allow us to extend measures to σ -algebras larger than the ones we start off with.

As an example of (b), consider the reals with Lebesgue measure: $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$. Every point is λ -null, since $\lambda\{x\} = \lambda[x, x] = x - x$. Hence the set \mathbb{Q} of rational numbers is λ -null: The set \mathbb{Q} is countable, and has an enumeration $\mathbb{Q} = \{q_n : n \in \mathbb{N}\}$. By countable additivity,

$$\lambda\mathbb{Q} = \sum_n \lambda\{q_n\} = 0$$

If the functions $f, g : \mathbb{R} \rightarrow \mathbb{R}$ are defined by

$$f = 0 \quad g = I_{\mathbb{Q}}$$

Then

$$f, g \text{ are equal } \lambda\text{-almost everywhere}$$

□

The following exercise is often useful:

Exercise 4.2.10 Suppose that $(\Omega, \mathcal{F}, \mu)$ is a measure space, and that $A \in \mathcal{F}$. Define

$$\mathcal{F} \cap A = \{F \cap A : F \in \mathcal{F}\}$$

(this is an *abuse of notation*), and let $\mu_A = \mu|_{\mathcal{F} \cap A}$. Then $(A, \mathcal{F} \cap A, \mu_A)$ is a measure space also — the restriction of $(\Omega, \mathcal{F}, \mu)$ to A .

□

4.3 Continuity Properties of Measures

4.3.1 Limit Operations on Sets

Definition 4.3.1 Let $(A_n)_{n \in \mathbb{N}}$ be a sequence of sets.

- We say that $(A_n)_n$ is increasing if $A_1 \subseteq A_2 \subseteq A_3 \subseteq \dots$. We say that $A_n \uparrow A$ if $(A_n)_n$ is increasing, and $\bigcup_n A_n = A$.
- Similarly, we say that $(A_n)_n$ is decreasing if $A_1 \supseteq A_2 \supseteq A_3 \supseteq \dots$. We say that $A_n \downarrow A$ if $(A_n)_n$ is decreasing, and $\bigcap_n A_n = A$.
- We define the *limit superior* of the sequence $(A_n)_n$ by

$$\limsup_n A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$$

- We define the *limit inferior* of the sequence $(A_n)_n$ by

$$\liminf_n A_n = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k$$

Also note the following simple interpretations of the above limit operations:

$$\begin{aligned} x \in \limsup_n A_n &\Leftrightarrow \forall n \left[x \in \bigcup_{k=n}^{\infty} A_k \right] \\ &\Leftrightarrow \forall n \exists k \geq n \left[x \in A_k \right] \\ &\Leftrightarrow x \text{ belongs to infinitely many of the sets } A_k \end{aligned}$$

Similarly,

$$\begin{aligned} x \in \liminf_n A_n &\Leftrightarrow \exists n \left[x \in \bigcap_{k=n}^{\infty} A_k \right] \\ &\Leftrightarrow \exists n \forall k \geq n \left[x \in A_k \right] \\ &\Leftrightarrow x \text{ belongs to all the } A_k \text{ from some } n \text{ onwards} \end{aligned}$$

In particular, $x \in \liminf_n A_n$ iff x belongs to all but finitely many of the A_n .

Thus we also write:

$$(A_n, \text{i.o.}) = \limsup_n A_n \quad (A_n, \text{ev.}) = \liminf_n A_n$$

where i.o. means *infinitely often*, and ev. means *eventually*.

Thus $x \in (A_n, \text{i.o.})$ iff x belongs to infinitely many of the sets A_n , etc.

Proposition 4.3.2 (a) $\liminf_n A_n \subseteq \limsup_n A_n$

(b) $(\limsup_n A_n)^c = \liminf_n A_n^c, \quad (\liminf_n A_n)^c = \limsup_n A_n^c$

Exercise 4.3.3 Prove Proposition 4.3.2.

□

4.3.2 Limits of Sets and Measures

Proposition 4.3.4 (Continuity properties)

Suppose that $(\Omega, \mathcal{F}, \mu)$ is a measure space, and let $A_1, A_2, \dots \in \mathcal{F}$.

(a) If $A_n \uparrow A$, then $\mu A_n \uparrow \mu A$.

(b) If $A_n \downarrow A$, and if $\mu A_k < \infty$ for some k , then $\mu A_n \downarrow \mu A$.

(c) If each $\mu A_k < \infty$, and $A_n \rightarrow A$, then $\mu A_n \rightarrow \mu A$.

Exercise 4.3.5 (a) Prove the preceding proposition.

Also show that (b) may fail if we drop the assumption that at least one of the μA_k is finite.

(b) Suppose that μ is *finitely additive* on the measurable space (Ω, \mathcal{F}) . Show that if

$$\mu A_n \rightarrow 0 \quad \text{whenever} \quad A_n \downarrow \emptyset$$

then μ is countably additive.

□

We end this section with some important results:

Proposition 4.3.6 (a) *FATOU'S LEMMA*: If μ is a measure on (Ω, \mathcal{F}) , and if $A_1, A_2, \dots \in \mathcal{F}$, then

$$\mu(\liminf_n A_n) \leq \liminf_n \mu A_n \quad \text{when} \quad \lim_n \mu A_n \text{ exists.}$$

(b) *REVERSE FATOU LEMMA*: If μ is a finite measure on (Ω, \mathcal{F}) , and if $A_1, A_2, \dots \in \mathcal{F}$, then

$$\mu(\limsup_n A_n) \geq \limsup_n \mu A_n \quad \text{when} \quad \lim_n \mu A_n \text{ exists.}$$

Proof: (a) Let $B_n = \bigcap_{m \geq n} A_m$. Then $\liminf_n A_n = \bigcup_n \bigcap_{m \geq n} A_m = \bigcup_n B_n$, so $B_n \uparrow \liminf_n A_n$. By Propn. 4.3.4(a), we see that $\mu B_n \uparrow \mu(\liminf_n A_n)$. Also, clearly $B_n \subseteq A_n$, and so $\mu A_n \geq \mu B_n$. It follows that

$$\liminf_n \mu A_n \geq \liminf_n \mu B_n = \mu(\liminf_n A_n)$$

(b) is left as an exercise. □

Exercise 4.3.7 We prove the reverse Fatou lemma:

- (a) Let $B_n = \bigcup_{m \geq n} A_m$. Explain why $B_n \downarrow \limsup_n A_n$. Conclude that $\mu B_n \downarrow \mu(\limsup_n A_n)$
- (b) Explain why $\mu B_n \geq \sup_{m \geq n} \mu A_m$.
- (c) Conclude that $\mu(\limsup_n A_n) \geq \limsup_{n \rightarrow \infty} \sup_{m \geq n} \mu A_m = \limsup_n \mu A_n$.
- (d) Where did we need the fact that μ is a finite measure?

□

Example 4.3.8 * Recall that the *Cantor set* is a subset of $[0, 1]$ which is constructed as follows: Let $C_0 = [0, 1]$. It is a single interval of length 1, so $\lambda(C_0) = 1$. Now let C_1 be C_0 with its *middle third* removed, i.e. $C_1 = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$. Thus C_1 consists of two disjoint intervals, each of length $\frac{1}{3}$. Hence $\lambda(C_1) = \frac{2}{3}$. Now remove the middle thirds of these two intervals to form C_2 , i.e. $C_2 = [0, \frac{1}{9}] \cup [\frac{2}{9}, \frac{1}{3}] \cup [\frac{2}{3}, \frac{7}{9}] \cup [\frac{8}{9}, 1]$. Then C_2 is a disjoint union of 4 intervals, each of length $\frac{1}{9}$, so $\lambda(C_2) = \frac{4}{9} = (\frac{2}{3})^2$. Continue in this way, removing the middle thirds of each of the intervals comprising C_n to form C_{n+1} . It follows that C_n consists of 2^n intervals, each of length $(\frac{1}{3})^n$, and thus that $\lambda(C_n) = (\frac{2}{3})^n$. Finally, let $C = \bigcap_{n=0}^{\infty} C_n$. C is the Cantor set.

The first thing to note is that C is a Borel set (why?)

The second thing to note is that C is not empty; in fact, C is uncountable. Here is one way to see this: Every real number $a \in [0, 1]$ can be written as an infinite sum $\sum_{i=1}^{\infty} \frac{a_i}{3^i}$, where $a_i = 0, 1$ or 2 . Thus the *ternary expansion* (as opposed to *decimal expansion*) of a is $0.a_1a_2a_3\dots$. For example, $\frac{1}{3} = 0.1000\dots$, $\frac{5}{9} = \frac{1}{3} + \frac{2}{9} = 0.1200\dots$, etc. A little thought will reveal that the Cantor set is formed by removing all numbers which have a 1 occurring in their ternary expansion. Thus C_1 is formed by removing all numbers which have a 1 in the first "decimal" place, C_2 is formed by removing all numbers in C_1 which have a 1 in the second "decimal" place, and so on. Thus the Cantor set is just the set of all numbers a in $[0, 1]$ which can be written as a sum $\sum_{i=1}^{\infty} \frac{a_i}{3^i}$, where $a_i = 0$ or 2 . Clearly there are uncountably many such.

Suppose now that once again we perform the random experiment of choosing a number from the interval $[0, 1]$, assuming as before that each number is equally likely to be chosen. The Lebesgue measure λ is the probability measure which models this situation, and we've stated (but not yet proved) that this measure can be defined on all Borel subsets of $[0, 1]$. We now ask: What is the probability that the number chosen belongs to C , i.e. what is $\lambda(C)$? Since $C \subseteq C_n$ for each $n \in \mathbb{N}$, and since $\lambda(C_n) = (\frac{2}{3})^n$, we must have $\lambda(C) = 0$. Thus the probability that the number chosen belongs to C is zero. Thus C , though a "large" set from the cardinality point of view, is a small set from the measure point of view. □

4.4 Lebesgue Measure from Coin Tossing

Consider again the random experiment of tossing a coin infinitely many times. We want to find an appropriate probability space for this experiment. It is clear that, if the coin is fair, each outcome is equally likely, and thus that the probability of a given outcome must be zero

(because there are infinitely many outcomes). We can say this before we have even decided upon an appropriate sample space. This we tackle now:

Letting 1 stand for “Heads” and 0 for “Tails”, we take the sample space to be

$$\hat{\Omega} = \{0, 1\}^{\mathbb{N}}$$

i.e. $\hat{\Omega}$ is the set of \mathbb{N} -indexed sequences of 0’s and 1’s. We take a slightly different view, however. Every sequence of 0’s and 1’s can be regarded as the *dyadic* or *binary expansion* of a real number. For example, the sequence

$$1 \ 1 \ 0 \ 1 \ 0 \ 0 \ 0 \ 1 \ \dots$$

can be thought of as the binary number

$$\begin{aligned} \underbrace{0.11010001\dots}_{\text{binary}} &= 1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2^2} + 0 \cdot \frac{1}{2^3} + 1 \cdot \frac{1}{2^4} + 0 \cdot \frac{1}{2^5} + 0 \cdot \frac{1}{2^6} + 0 \cdot \frac{1}{2^7} + 1 \cdot \frac{1}{2^8} + \dots \\ &= \underbrace{0.816\dots}_{\text{decimal}} \end{aligned}$$

We thus have a correspondence between sequences $(a_n : n \in \mathbb{N})$ of 0’s and 1’s, and real numbers between 0 and 1:

$$(a_n : n \in \mathbb{N}) \mapsto \sum_{n=1}^{\infty} \frac{a_n}{2^n}$$

Clearly $(0, 0, 0, \dots) \mapsto 0$, $(1, 1, 1, \dots) \mapsto 1$, $(1, 0, 0, 0, \dots) \mapsto \frac{1}{2}$, etc. In this way, every sequence of 0’s and 1’s provides us with a unique number between 0 and 1. The only problem is that this correspondence is not one-to-one:

$$(1, 0, 0, 0, \dots) \mapsto \frac{1}{2} \quad (0, 1, 1, 1, \dots) \mapsto \frac{1}{2}$$

However, we can deal with this in a clever way. Call a dyadic expansion *terminating* if it eventually ends in all 0’s (tails), i.e. if there are only finitely many 1’s. Let $T = \{\omega \in \hat{\Omega} : \omega \text{ is terminating}\}$. It is clear that every terminating expansion has associated with it a non-terminating expansion that corresponds to the same number: If $a_l = 1$ is the *last* 1 in the terminating sequence $(a_n : n \in \mathbb{N})$

$$\underbrace{0.a_1a_2a_3\dots a_{l-1}a_l000\dots}_{\text{binary}} = \underbrace{0.a_1a_2a_3\dots a_{l-1}0111\dots}_{\text{binary}}$$

Moreover, a little thought shows that if we chuck out the terminating sequences from $\hat{\Omega}$, then the correspondence above is a bijection between $\hat{\Omega} - T$ and $(0, 1]$. How many such terminating dyadic expansions are there? It is not hard to see that T is countable. Moreover, since each element of $\hat{\Omega}$ has probability 0, the probability of the event T is also 0 (being a countable union of events of probability 0). It is therefore *practically certain* that the event T won’t occur. The *terminating* dyadic expansions are therefore, in a sense, redundant. Nothing is lost by chucking them out (except a set of measure 0). We may therefore take the sample space to be the set

$$\Omega = (0, 1]$$

A natural algebra for this experiment is

$$\mathcal{F} = \text{all events that can be decided after only finitely many tosses}$$

Here are some examples of elements of \mathcal{F} :

- (a) A = the first toss is “Heads”. These are the dyadic numbers with a 1 in the first place, i.e. all the numbers from $0.1000\dots = \frac{1}{2}$ to $0.1111\dots = 1$, but not including $0.1000\dots$, because it is terminating. Thus $A = (\frac{1}{2}, 1]$
- (b) B = the third toss is “Tails”. This is the set of all dyadic numbers with a 0 in the third place. These are the numbers in intervals

binary	=	decimal
$(0.000000\dots, 0.000111\dots]$	$=$	$(0, 0.125]$
$(0.010000\dots, 0.010111\dots]$	$=$	$(0.25, 0.375]$
$(0.100000\dots, 0.100111\dots]$	$=$	$(0.5, 0.625]$
$(0.110000\dots, 0.110111\dots]$	$=$	$(0.75, 0.875]$

$$\text{Hence } B = (0, 0.125] \cup (0.25, 0.375] \cup (0.5, 0.625] \cup (0.75, 0.875].$$

- (c) C = There are 2 “Heads” and 1 “Tail” in the first 3 tosses. A little thought shows that

$$C = (0.375, 0.5] \cup (0.625, 0.75] \cup (0.75, 0.875]$$

Reasoning along these lines, it is clear that \mathcal{F} is the σ -algebra generated by all intervals of the form $(\frac{k}{2^n}, \frac{k+1}{2^n}]$, where $n \in \mathbb{N}$ and $k < 2^n$. It is therefore not hard to see that $\mathcal{F} = \mathcal{B}(0, 1]$ (since every real number can be approximated arbitrarily closely by a dyadic rational, i.e. a number of the form $\frac{k}{2^n}$).

Having identified the “right” σ -algebra, we turn to the probability measure appropriate for this experiment.

- (a) $\mathbb{P}(A)$ = probability that first toss is “Heads”. Assuming a fair coin, this is clearly $\frac{1}{2}$. Now the event A corresponds to the interval $(\frac{1}{2}, 1]$, and $\lambda(\frac{1}{2}, 1] = \frac{1}{2}$ (where λ is the *Lebesgue measure* introduced in Example 4.2.2. Thus $\mathbb{P}(A) = \lambda(A)$.
- (b) $\mathbb{P}(B)$ = probability that third toss lands “Tails”. This is clearly also $\frac{1}{2}$, as the third toss is just as likely to land “Heads” as it is “Tails”. Now $\lambda(B) = \lambda(0, 0.125] + \lambda(0.25, 0.375] + \lambda(0.5, 0.625] + \lambda(0.75, 0.875] = 4 \times \frac{1}{8} = \frac{1}{2}$, and thus $\mathbb{P}(B) = \lambda(B)$ in this case also.
- (c) $\mathbb{P}(C)$ = probability that there are 2 heads and 1 tail in the first 3 tosses. This probability is clearly $\binom{3}{2}2^{-3} = \frac{3}{8}$. In this case we therefore also have $\mathbb{P}(C) = \lambda(C)$.

It therefore becomes apparent that the “right” probability space for the random experiment of tossing a coin infinitely many times is just the *same* as the random experiment of picking a number from $(0, 1]$ (Example 1.3.4).

All of this leads us to formulate the following principle:

Borel's Principle:

Consider the random experiment of tossing a (fair) coin infinitely many times, and let E be an event. Interpret E as a subset of $(0, 1]$. Then $\mathbb{P}(E) = \lambda(E)$, i.e. the probability that the event E occurs is just the Lebesgue measure of the associated subset of the unit interval.

Exercise 4.4.1 This exercise is meant to get you thinking. Don't worry too much about the details.

- (a) Consider a random experiment in which a coin is tossed 3 times. Describe a suitable sample space and σ -algebra for this experiment. How many elements does this σ -algebra have?
- (b) Consider a random experiment in which a coin is tossed infinitely many times. From lecture notes, you know that we can take the sample space to be the interval $(0, 1]$.
 - (i) What is the smallest σ -algebra on $(0, 1]$ which contains the information about the outcomes of the first 3 coin tosses? How many elements does it have?
 - (ii) Compare your answer in (i) to that in (a).
 - (iii) What is the smallest σ -algebra on $(0, 1]$ which contains information about the outcomes of all the tosses? [To know the outcome of all the tosses, you have to know the outcome of the first n tosses for every $n \in \mathbb{N}$.]

□

Exercise 4.4.2 Consider the experiment consisting of an infinite sequence of coin flips. For each of the following, find the Borel set that corresponds to the event, and calculate its probability.

- (a) The first head comes immediately after an even number of tails.
- (b) At six flips, but no earlier, the number of heads equals the number of tails.
- (c) The sequence HHT occurs before the sequence THT.

□

Exercise 4.4.3 A gambler starts with an initial stake of 2 Rands. She bets at even odds on a coin toss, i.e. she wins one rand if the coin lands H, and loses one rand if it lands T. We are going to investigate the event that she is *ruined* i.e. loses all her stake.

If she keeps playing until she is ruined, this may involve a potentially infinite number of coin tosses. From lectures, we know that this situation may be modelled by the probability space $((0, 1], \mathcal{B}(0, 1], \lambda)$, where λ is the Lebesgue measure on $(0, 1]$. The gambler's sequence of play is described by an outcome $\omega = 0.\omega_1\omega_2\omega_3\dots$ which is a sequence of 0's and 1's representing a real number in binary.

- (a) What is the probability that she is ruined on the 1st toss?
- (b)
 - (i) What is the probability that she is ruined on the 2nd toss?
 - (ii) Which Borel set corresponds to her being ruined on the 2nd toss?
 - (iii) What is the Lebesgue measure of this set?
- (c) What is the probability that she is ruined on the 3rd toss (but not before)?
- (d)
 - (i) What is the probability that she is ruined on the 4th toss, but not before?
 - (ii) Which Borel set corresponds to her being ruined on the 4th toss?
 - (iii) What is the Lebesgue measure of this set?

We now investigate a little deeper: Let $s_n(\omega)$ be the amount that she has won (or lost, if negative) after n tosses if the outcome is ω . Thus $s_n(\omega) = \sum_{k=1}^n r_k(\omega)$ where the $r_k(\omega)$ are the *Rademacher functions*, defined by

$$r_k(\omega) = \begin{cases} 1 & \text{if } \omega_k = 1 \text{ i.e. } k^{\text{th}} \text{ toss lands Heads} \\ -1 & \text{if } \omega_k = 0 \text{ i.e. } k^{\text{th}} \text{ toss lands Tails} \end{cases}$$

- (e) Let N be an integer. Explain why $\{\omega \in (0, 1] : s_n(\omega) = N\}$ is a Borel set. What is the measure of this set?
- (f) Let N be an integer. Explain why $\{\omega \in (0, 1] : s_n(\omega) > N\}$ is a Borel set. Conclude that each s_n is a measurable function (i.e. a random variable).
- (g) Explain why the set

$$R_n = \{\omega \in (0, 1] : s_k(\omega) > -2 \text{ for } 1 \leq k < n, s_n(\omega) = -2\}$$

is a Borel set. What event does it describe?

- (h) What event is described by the set $R = \bigcup_{n=1}^{\infty} R_n$?

□

4.5 Some Probability Theory

Probability is all about *information*. I toss a coin and see that it lands heads. You don't see the coin. For you the probability that the coin has landed heads is $\frac{1}{2}$, but for me it is 1. New information changes the probability measures.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let A, B be events. Knowledge that B has occurred can change our estimation of the probability that A has occurred. We write $\mathbb{P}(A|B)$ for the probability that A occurs given that we know that B has occurred. We call $\mathbb{P}(A|B)$ the **conditional probability** of A given B .

Example 4.5.1 A die is rolled. Let A be the event that the outcome is a 6, let B be the event that the outcome is an even number, and let C be the event that the outcome is an odd number. Clearly $\mathbb{P}(A) = \frac{1}{6}$. However, if we know for sure that the outcome is an even number, then the probability of getting a 6 is $\frac{1}{3}$, i.e. $\mathbb{P}(A|B) = \frac{1}{3}$. In the same way, if B occurs, then C cannot possibly occur, so although $\mathbb{P}(C) = \frac{1}{2}$, $\mathbb{P}(C|B) = 0$.

□

Basically, what's happening here is that we have to *modify our probability measure* to accommodate the “new” information that B has occurred. If $\mathbb{P}(\cdot|B)$ is the new probability measure on (Ω, \mathcal{F}) , then we must have $\mathbb{P}(B|B) = 1$ and $\mathbb{P}(\Omega - B|B) = 0$. If A is another event, then A occurs if and only if $A \cap B$ occurs, since we *know* that B also occurs, and it makes sense to assume that the new probability that A occurs is proportional to the old probability that $A \cap B$ occurs, i.e. that $\mathbb{P}(A|B) = c\mathbb{P}(A \cap B)$ for some constant c . To ensure $\mathbb{P}(B|B) = 1$, we must have $c = \mathbb{P}(B)^{-1}$. We therefore find that

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

the standard formula given in elementary probability theory texts.

Exercises 4.5.2 (1.) Prove that $\mathbb{P}(\cdot|B)$ is a probability measure on (Ω, \mathcal{F}) .

(2.) For events A_1, \dots, A_n , prove that

$$\mathbb{P}(A_1 \cap \dots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_1 \cap A_2) \dots \mathbb{P}(A_n|A_1 \cap \dots \cap A_{n-1})$$

□

Example 4.5.3 A couple has two children. Assuming that boys and girls are equally likely, and given that one of the children is a girl, what is the probability that the other child is also a girl?

We can model this probability space as follows:

$$\begin{aligned}\Omega &= \{BB, BG, GB, GG\} \\ \mathcal{F} &= \mathcal{P}(\Omega) \\ \forall \omega \in \Omega [\mathbb{P}(\omega) &= \frac{1}{4}]\end{aligned}$$

Let B be the event that at least one of the children is a girl, i.e. $B = \{GB, BG, GG\}$, and let A be the event that both children are girls, i.e. $A = \{GG\}$. Then

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{1/4}{3/4} = \frac{1}{3}$$

□

Two events A, B are said to be *independent* if knowledge of B tells us nothing about A , and vice versa. By this we mean that our estimation of the probability that A occurs isn't changed by the knowledge that B has occurred. Thus:

$$\mathbb{P}(A) = \mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

and hence we have the multiplication law

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$$

The above equation gives us the definition of independent events:

Definition 4.5.4 Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A (possibly infinite) set $\mathcal{A} = \{A_i : i \in I\}$ of events is said to be an **independent family** provided that for any *distinct* $i_1, i_2, \dots, i_n \in I$

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_n}) = \mathbb{P}(A_{i_1})\mathbb{P}(A_{i_2}) \dots \mathbb{P}(A_{i_n})$$

Example 4.5.5 (a) Consider the random trial of tossing coin twice. The sample space Ω is the 4-element set $\{HH, HT, TH, TT\}$ and the associated σ -algebra is just $\mathcal{P}(\Omega)$. Intuitively, if the coin is fair, the outcome of the first coin should have no influence on the second. Thus knowing that the first coin has landed heads should make no difference to whether the second coin lands heads. Let $B = \{HH, HT\}$ be the event that the first coin lands heads, and let $A = \{HH, TH\}$ be the event that the second coin lands heads. Then $\mathbb{P}(A \cap B) = \mathbb{P}(\{HH\}) = \frac{1}{4}$, and $\mathbb{P}(A) \cdot \mathbb{P}(B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. Thus $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$, i.e. the events A and B are indeed independent.

(b) Consider the same experiment as in (a), but with one important difference: Before the experiment starts, we are told that the coin is *unfair*. It has either two heads, or two tails, but we are not told which. Each possibility is equally likely.

To model this, we use a different probability measure \mathbb{Q} , which has

$$\mathbb{Q}(\{HH\}) = \frac{1}{2} = \mathbb{Q}(\{TT\}) \quad \mathbb{Q}(\{HT\}) = 0 = \mathbb{Q}(\{TH\})$$

In this case $\mathbb{Q}(A \cap B) = \frac{1}{2}$, whereas $\mathbb{Q}(A)\mathbb{Q}(B) = \frac{1}{4}$. Thus A and B are not independent under \mathbb{Q} .

□

It's worth pointing out once more that it depends on the probability measure whether or not two events are independent, i.e. it is possible for events to be independent under one measure, but not under another. The notion of independence is therefore a probabilistic notion, which has no analogue in general measure theory.

The intuitive idea about independence was the following: Two events are independent if the information that one of the events has occurred does not tell us anything new about the other, i.e. it does not lead us to revise our estimation of its probability. Now σ -algebras are the carriers of information, and we would therefore like a definition of independence which involves σ -algebras. We therefore define independence anew:

Definition 4.5.6 Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Sub- σ -algebras $\mathcal{G}_1, \mathcal{G}_2, \dots$ of \mathcal{F} are said to be *independent* if events in *distinct* \mathcal{G}_n are independent, i.e. whenever $n_1, n_2, \dots, n_m \in \mathbb{N}$ are *distinct* positive integers and $G_{n_1} \in \mathcal{G}_{n_1}, G_{n_2} \in \mathcal{G}_{n_2}, \dots, G_{n_m} \in \mathcal{G}_{n_m}$ are events, we have

$$\mathbb{P}(G_{n_1} \cap G_{n_2} \cap \dots \cap G_{n_m}) = \prod_{k=1}^m \mathbb{P}(G_{n_k})$$

The basic idea is that two σ -algebras are independent if there is no information about an event in one of the σ -algebras that would lead us to revise our estimate of the probability of any event in the other σ -algebra.

Example 4.5.7 Suppose that A, B are events in some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then $\mathcal{A} = \{\Omega, A, A^c, \emptyset\}$ and $\mathcal{B} = \{\Omega, B, B^c, \emptyset\}$ are the σ -algebras of events that can be decided by knowledge of A, B respectively. It is easy to show that A, B are independent events if and only if \mathcal{A}, \mathcal{B} are independent σ -algebras. For example, $\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c)$, and thus $\mathbb{P}(A \cap B^c) = \mathbb{P}(A)[1 - \mathbb{P}(B)] = \mathbb{P}(A)\mathbb{P}(B^c)$, by independence of A, B . It follows that A, B^c are independent if A, B are. The other combinations of events are similarly proven independent.

□

Remarks 4.5.8 Can an event be independent of itself, i.e. given an event A , can the events A, A be independent? Here we have to be a little careful. From the intuitive point of view, the answer would seem to be no, since the information that the event A has occurred will certainly make us re-evaluate our estimation of the probability that A has occurred! However, if we look at the definition, A and A will be independent provided $\mathbb{P}(A \cap A) = \mathbb{P}(A)\mathbb{P}(A)$, i.e. provided $\mathbb{P}(A) = \mathbb{P}(A)^2$. This can happen only if $\mathbb{P}(A)$ is either 0 or 1. That's not too far removed from our intuition. If $\mathbb{P}(A) = 1$, for example, then A happens almost surely, so telling us that A has happened does not really give us any information. We were practically certain that it would anyway.

□

Exercises 4.5.9 (1.) A gambling game involves the rolling of a fair die followed by the flipping of a fair coin.

- (a) Set up a reasonable probability space to model this situation.
 - (b) Let A be the event that the die lands on an even number, and let B be the event that the coin lands tails. Show that A and B are independent events.
- (2.) An HIV-test is 95% accurate, i.e. it gives the correct result 95% of the time. John lives in a small town with a 1000 inhabitants, 50 of whom have AIDS. After a particularly wild night, John decides to have an HIV-test, which comes out positive. What is the probability that John has AIDS?

□

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $\mathcal{A} = \{A_n : n \in \mathbb{N}\}$ be a countable set of events. It may help to think of \mathcal{A} as a sequence of events, A_{n+1} following A_n .

Proposition 4.5.10 (First Borel–Cantelli Lemma)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $\{A_n : n \in \mathbb{N}\}$ be a sequence of events. If

$$\sum_n \mathbb{P}(A_n) < \infty$$

then

$$\mathbb{P}(A_n, \text{i.o.}) = 0$$

Proof: Let $B_n = \bigcup_{k=n}^{\infty} A_k$. Then $B_n \downarrow \limsup A_n = (A_n, \text{i.o.})$. Hence

$$\mathbb{P}(A_n, \text{i.o.}) \leq \mathbb{P}(B_n) \leq \sum_{k=n}^{\infty} \mathbb{P}(A_k)$$

for all $n \in \mathbb{N}$. Now as $n \rightarrow \infty$, the right-hand sum goes to zero, since $\sum_n \mathbb{P}(A_n)$ converges. Hence $\mathbb{P}(A_n, \text{i.o.}) = 0$. +

Proposition 4.5.11 (Second Borel–Cantelli Lemma)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $\{A_n : n \in \mathbb{N}\}$ be a family of independent events. If

$$\sum_n \mathbb{P}(A_n) = \infty$$

then

$$\mathbb{P}(A_n, \text{i.o.}) = 1$$

Proof: The proof depends on the fact that $1 - x \leq e^{-x}$ for all $x \in \mathbb{R}$, an inequality which is easily proved using first-year calculus. Now clearly $(A_n, \text{i.o.}) = \limsup A_n = (\liminf A_n^c)^c = (A_n^c, \text{ev.})^c$, and thus it suffices to prove that $\mathbb{P}(A_n^c, \text{ev.}) = 0$. But $(A_n^c, \text{ev.}) = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k^c$ by definition, and so it suffices to show that $\mathbb{P}(\bigcap_{k=n}^{\infty} A_k^c) = 0$ for all n . Now by independence of the A_n , and thus the A_n^c , we have

$$\mathbb{P}\left(\bigcap_{k=n}^{n+m} A_k^c\right) = \prod_{k=n}^{n+m} [1 - \mathbb{P}(A_k)] \leq \prod_{k=n}^{n+m} e^{-\mathbb{P}(A_k)} = e^{-\sum_{k=n}^{n+m} \mathbb{P}(A_k)}$$

for all $m \in \mathbb{N}$. Now since $\sum_n \mathbb{P}(A_n)$ diverges, the power of e on the right must tend to zero as $m \rightarrow \infty$. Thus $\mathbb{P}(\bigcap_{k=n}^{\infty} A_k^c) = \lim_{m \rightarrow \infty} \mathbb{P}(\bigcap_{k=n}^{n+m} A_k^c) = 0$, as required. +

Remarks 4.5.12 The First Borel–Cantelli Lemma says that given events A_n , *not necessarily independent*, if the sum of the probabilities $\mathbb{P}(A_n)$ converges, then $(A_n, \text{i.o.})$ is an event of zero probability. The Second Borel–Cantelli Lemma says that if the A_n are independent and the sum of the probabilities $\mathbb{P}(A_n)$ diverges, then the event $(A_n, \text{i.o.})$ occurs almost surely, i.e. with probability 1. Thus for independent events A_n , there is no middle road: $(A_n, \text{i.o.})$ is either an event of probability 0 or an event of probability 1.

□

Exercise 4.5.13 It is sometimes asserted that if a monkey hits the keys of a type writer at random, it would eventually produce, in one continuous stream, the complete works of William Shakespeare. Prove it.

□

4.6 Extension of Measures*

4.6.1 Other Families of Sets*

It turns out that σ -algebras can be quite complicated to deal with, especially if the sample space is finite. In many cases, it is easier to work with simpler classes of sets, especially π -systems.

Definition 4.6.1 Let \mathcal{C} be a collection of subsets of Ω

- (a) \mathcal{C} is called a π -system if it is closed under finite intersections.
- (b) \mathcal{C} is called a λ -system if
 - (i) $\Omega \in \mathcal{C}$;
 - (ii) $A, B \in \mathcal{C}$ and $A \subseteq B$ implies $B - A \in \mathcal{C}$;
 - (iii) If $A_1, A_2, \dots \in \mathcal{C}$ and $A_n \uparrow A$, then $A \in \mathcal{C}$.
- (c) We denote by $\pi(\mathcal{C})$ and $\lambda(\mathcal{C})$ the π -, respectively, λ -system *generated* by \mathcal{C} , i.e. the smallest π -, respectively, λ -system on Ω which contains \mathcal{C} .

Why do $\pi(\mathcal{C})$, $\lambda(\mathcal{C})$ always exist? It follows from the easily proved fact that the intersection of an arbitrary family of π -systems (resp. λ -systems) is again a π -system (resp. λ -system).

Proposition 4.6.2 A family \mathcal{C} of subsets of Ω is a σ -algebra iff it is both a π -system and a λ -system.

Proof: It is clear that a σ -algebra is also a π -system and a λ -system.

Conversely, suppose \mathcal{C} is both a π - and a λ -system. Then \mathcal{C} is closed under complementation, by (i), (ii) of Defn. 8.1.1(b). De Morgan's Laws applied to Defn. 8.1.1(a) show that \mathcal{C} is closed under *finite* unions. Finally, given $A_1, A_2, \dots \in \mathcal{C}$, let $A = \bigcup_n A_n$. Define

$$B_n = \bigcup_{m \leq n} A_m$$

Then each $B_n \in \mathcal{C}$, and $B_n \uparrow A$. Hence $A \in \mathcal{C}$, by Defn. 8.1.1(b)(iii), and thus \mathcal{C} is closed under countable unions.

◄

The following technical result often allows us to work with “easy” π -systems, instead of the “difficult” σ -algebras:

Theorem 4.6.3 (Dynkin’s Lemma, Monotone Class Theorem)

(a) If \mathcal{C} is a π -system on Ω , then

$$\lambda(\mathcal{C}) = \sigma(\mathcal{C})$$

(b) Suppose that \mathcal{C} is a π -system and that \mathcal{D} is a λ -system (both on a set Ω), and also that $\mathcal{C} \subseteq \mathcal{D}$. Then $\sigma(\mathcal{C}) \subseteq \mathcal{D}$.

Proof: (a) Let $\mathcal{D} = \lambda(\mathcal{C})$. By Propn. 8.1.2, it suffices to show that \mathcal{D} is a π -system. We do this in two steps.

STEP I: Fix $C \in \mathcal{C}$, and define

$$\mathcal{D}_C = \{A \in \mathcal{D} : A \cap C \in \mathcal{D}\}$$

Then $\mathcal{C} \subseteq \mathcal{D}_C \subseteq \mathcal{D}$ (because \mathcal{C} is a π -system). We now show that $\mathcal{D}_C = \mathcal{D}$. To that end, it suffices to show that \mathcal{D}_C is a λ -system (because then \mathcal{D}_C is a λ -system containing \mathcal{C} , and \mathcal{D} is the *smallest* such). We therefore verify (i)-(iii) of Defn. 8.1.1:

(i) is obvious.

If $A, B \in \mathcal{D}_C$ and $A \subseteq B$, then $(B - A) \cap C = (B \cap C) - (A \cap C)$. But $B \cap C, A \cap C \in \mathcal{D}$ by definition of \mathcal{D}_C , and thus $(B - A) \cap C \in \mathcal{D}$, because \mathcal{D} is a λ -system. Thus $(B - A) \in \mathcal{D}_C$. Finally, if $A_1, A_2, \dots \in \mathcal{D}_C$ and $A_n \uparrow A$, then $A_1 \cap C, A_2 \cap C, \dots \in \mathcal{D}$ and $(A_n \cap C) \uparrow A \cap C$. Hence $A \cap C \in \mathcal{D}$, and so $A \in \mathcal{D}_C$.

We now know that $\mathcal{D}_C = \mathcal{D}$ for *every* $C \in \mathcal{C}$.

STEP II: Now, fix any $D \in \mathcal{D}$, and define

$$\mathcal{D}^D = \{A \in \mathcal{D} : A \cap D \in \mathcal{D}\}$$

First note that if $C \in \mathcal{C}$, then $\mathcal{D}_C = \mathcal{D}$, so $D \in \mathcal{D}_C$. It follows that $D \cap C \in \mathcal{D}$, and thus that $C \in \mathcal{D}^D$, for every $C \in \mathcal{C}$. Thus $\mathcal{C} \subseteq \mathcal{D}^D$, for all $D \in \mathcal{D}$.

It follows as above that \mathcal{D}^D is a λ -system, and thus that $\mathcal{D}^D = \mathcal{D}$, for all $D \in \mathcal{D}$.

In particular, if $A, B \in \mathcal{D}$, then $A \in \mathcal{D}^B$, and so $A \cap B \in \mathcal{D}$. This shows that \mathcal{D} is a π -system, and thus a σ -algebra.

(b) follows directly from (a). (Why?)

+

4.6.2 The Extension Theorem*

This section is mainly technical: The Carathéodory Extension Theorem (Thm. 4.6.9) ensures that commonly used measures, such as Lebesgue measure, *exist*. Before we tackle that, here is an easy but useful result: Two probability measures which agree on a π -system agree on the σ -algebra generated by that π -system.

Proposition 4.6.4 Suppose that μ_1, μ_2 are finite measures on a measurable space (Ω, \mathcal{F}) , and let \mathcal{C} be a π -system such that $\Omega \in \mathcal{C}$ and $\sigma(\mathcal{C}) = \mathcal{F}$. Then $\mu_1 = \mu_2$ iff μ_1, μ_2 agree on \mathcal{C} .

Exercise 4.6.5 Prove the preceding proposition.

[Hint: Show that $\mathcal{D} = \{A \in \mathcal{F} : \mu_1 A = \mu_2 A\}$ is a λ -system, and that $\mathcal{C} \subseteq \mathcal{D}$. Then apply Dynkin's Lemma (Thm. 8.1.3)]

□

Definition 4.6.6 Let Ω be a non-empty set.

(a) A map $\mu : \mathcal{P}(\Omega) \rightarrow [0, +\infty]$ is said to be an *outer measure* on Ω iff:

- (i) $\mu \emptyset = 0$;
- (ii) μ is *monotone*: $A \subseteq B$ implies $\mu A \leq \mu B$;
- (iii) μ is countably *sub-additive*: If $A_1, A_2, \dots \subseteq \Omega$, then $\mu(\bigcup_n A_n) \leq \sum_n \mu A_n$.

(b) A set $A \subseteq \Omega$ is said to be μ -*measurable* if and only if

$$\mu E = \mu(E \cap A) + \mu(E \cap A^c) \quad \text{for all } E \subseteq \Omega$$

Note that we require μA to be defined for every subset $A \subseteq \Omega$. We haven't mentioned a base σ -algebra. But there is one:

Theorem 4.6.7 Let μ be an outer measure on Ω and let $\mathcal{M}(\mu)$ be the family of all μ -measurable sets. Then $(\Omega, \mathcal{M}(\mu), \mu|_{\mathcal{M}(\mu)})$ is a measure space.

Proof: We must show that $\mathcal{M}(\mu)$ is a σ -algebra, and that μ is countably additive on $\mathcal{M}(\mu)$.

Certainly \emptyset is μ -measurable. Also, it is obvious that $A \in \mathcal{M}(\mu)$ implies $A^c \in \mathcal{M}(\mu)$.

Next, we show that $\mathcal{M}(\mu)$ is closed under finite intersections: Let $A, B \in \mathcal{M}(\mu)$, and let $E \subseteq \Omega$. Then

$$\begin{aligned} \mu E &= \mu(E \cap A) + \mu(E \cap A^c) \\ &= \mu(E \cap A \cap B) + \mu(E \cap A \cap B^c) + \mu(E \cap A^c) \\ &= \mu(E \cap (A \cap B)) + \mu(E \cap (A \cap B)^c) \\ &\geq \mu E \end{aligned}$$

where the third line follows because $\mu(E \cap (A \cap B)^c) = \mu(E \cap (A \cap B)^c \cap A) + \mu(E \cap (A \cap B)^c \cap A^c) = \mu(E \cap B \cap A^c) + \mu(E \cap A^c)$, and the final line holds because μ is sub-additive. It follows that $A \cap B \in \mathcal{M}(\mu)$. Hence $\mathcal{M}(\mu)$ is an algebra.

Next, let $A, B \in \mathcal{M}(\mu)$ be disjoint, and let $E \subseteq \Omega$. Then $B \subseteq A^c$, and hence

$$\mu(E \cap (A \cup B)) = \mu(E \cap (A \cup B) \cap A) + \mu(E \cap (A \cup B) \cap A^c) = \mu(E \cap A) + \mu(E \cap B) \quad (*)$$

Specializing to $E = \Omega$ proves that μ is finitely additive on $\mathcal{M}(\mu)$.

Now let A_1, A_2, \dots be a countable sequence of disjoint elements of $\mathcal{M}(\mu)$, and let $E \subseteq \Omega$. Define $B_n = \bigcup_{m \leq n} A_m$, and let $A = \bigcup_n A_n = \bigcup_n B_n$. Then by monotonicity and (*), we see that

$$\mu(E \cap A) \geq \mu(E \cap B_n) = \sum_{m \leq n} \mu(E \cap A_m)$$

As $n \rightarrow \infty$, and invoking the fact that μ is subadditive, we see that

$$\mu(E \cap A) \geq \sum_n \mu A_n \geq \mu(A) \geq \mu(E \cap A) \quad (**)$$

so that equality holds throughout. Countable additivity of μ on $\mathcal{M}(\mu^*)$ is then obtained by specializing to the case $E = \Omega$.

Also, since $B_n \in \mathcal{M}(\mu)$, we see, by monotonicity and (**), that

$$\begin{aligned} \mu E &= \mu(E \cap B_n) + \mu(E \cap B_n^c) \\ &\geq \sum_{m \leq n} \mu(E \cap A_m) + \mu(E \cap A^c) \\ &\rightarrow \mu(E \cap A) + \mu(E \cap A^c) \\ &\geq \mu E \end{aligned}$$

Hence equality holds throughout, and so $A \in \mathcal{M}(\mu)$. This proves that $\mathcal{M}(\mu)$ is a σ -algebra.

⊥

Here is one of the most important ways of obtaining outer measures:

Proposition 4.6.8 *Let \mathcal{A} be an algebra on a set Ω , and let μ_0 be a non-negative countably additive set function on \mathcal{A} . Define $\mu^* : \mathcal{P}(\Omega) \rightarrow [0, +\infty]$ by*

$$\mu^* E = \inf \left\{ \sum_n \mu_0 A_n : (A_n)_n \text{ a sequence of sets in } \mathcal{A} \text{ with } E \subseteq \bigcup_n A_n \right\}$$

Then μ^ is an outer measure on Ω which extends μ_0 .*

Moreover, $\mathcal{A} \subseteq \mathcal{M}(\mu^)$.*

Proof: To see that μ^* extends μ_0 is easy: For let $A \in \mathcal{A}$. Firstly, if $(A_n)_n$ is a sequence in \mathcal{A} which covers A , then $\mu_0 A \leq \sum_n \mu_0 A_n$ (by countable additivity of μ_0), and thus $\mu_0 A \leq \mu^* A$. The sequence $(A, \emptyset, \emptyset, \dots)$ witnesses the fact that $\mu^* A \leq \mu_0 A$ as well.

Next, we prove that μ^* is an outer measure on Ω .

- (i) $\mu^* \emptyset = \mu_0 \emptyset = 0$;
- (ii) If $E \subseteq F$, then any \mathcal{A} -covering of F is also a covering of E , and thus $\mu^* E \leq \mu^* F$.
- (iii) Finally, suppose that $E_n \subseteq \Omega$, and let $E = \bigcup_n E_n$. We must show that

$$\mu^* E \leq \sum_n \mu^* E_n \tag{*}$$

Fix $\varepsilon > 0$. For each E_n , choose a sequence $(A_{n,k})_k$ in \mathcal{A} such that

$$\sum_k \mu_0 A_{n,k} \leq \mu^* E_n + \varepsilon 2^{-n}$$

Then $\{A_{n,k} : n, k \in \mathbb{N}\}$ is a countable \mathcal{A} -covering of E , and thus

$$\mu^* E \leq \sum_{n,k} \mu_0 A_{n,k} \leq \sum_n \left(\sum_k \mu_0 A_{n,k} \right) = \sum_n (\mu^* E_n + \varepsilon 2^{-n}) = \sum_n \mu^* E_n + \varepsilon$$

Since ε is arbitrary (> 0), (*) follows.

Next, we show that every member of \mathcal{A} is μ^* -measurable. So let $B \in \mathcal{M}(\mu^*)$, and let $E \subseteq \Omega$ be arbitrary. Let $\varepsilon > 0$. Choose a sequence $(A_n)_n$ in \mathcal{A} which covers E , and such that

$$\sum_n \mu_0 A_n \leq \mu^* E + \varepsilon$$

Then $(A_n \cap B)_n, (A_n \cap B^c)_n$ are, respectively, sequences in \mathcal{A} which cover $E \cap B$ and $E \cap B^c$. It follows that

$$\mu^*(E \cap B) + \mu^*(E \cap B^c) \leq \sum_n \left(\mu_0(A_n \cap B) + \mu_0(A_n \cap B^c) \right) = \sum_n \mu^* A_n \leq \mu^* E + \varepsilon$$

+

Since $\varepsilon > 0$ is arbitrary, it follows that

$$\mu^*(E \cap B) + \mu^*(E \cap B^c) = \mu^* E \quad \text{for all } E \subseteq \Omega$$

(we proved \leq , but \geq always holds, by sub-additivity.) Hence B is μ^* -measurable, for every $B \in \mathcal{A}$.

+

Theorem 4.6.9 (Carathéodory's Extension Theorem)

Let \mathcal{A} be an algebra on a set Ω , and let μ_0 be a countably additive non-negative set function^a on \mathcal{A} , such that $\mu_0 \emptyset = 0$. Then μ_0 extends to a measure μ on $\sigma(\mathcal{A})$, i.e. there is a measure μ on $(\Omega, \sigma(\mathcal{A}))$ such that $\mu_0 A = \mu A$ for all $A \in \mathcal{A}$.

Moreover, if μ_0 is σ -finite, then the extension μ of μ_0 is unique.

^aThis means that if $A_1, A_2, \dots \in \mathcal{A}$ are mutually disjoint, and if also $\bigcup_n A_n \in \mathcal{A}$, then $\mu_0(\bigcup_n A_n) = \sum_n \mu_0 A_n$.

Proof: Let $\mathcal{F} = \sigma(\mathcal{A})$, and define μ^* on $(\Omega, \mathcal{P}(\Omega))$ by

$$\mu^* E = \inf \left\{ \sum_n \mu_0 A_n : (A_n)_n \text{ a sequence of sets in } \mathcal{A} \text{ with } E \subseteq \bigcup_n A_n \right\}$$

Then by Propn. 4.6.8, μ^* is an outer measure which extends μ_0 . By Theorem 4.6.7, μ^* is a measure on the σ -algebra $\mathcal{M}(\mu^*)$ of all μ^* -measurable sets, and $\mathcal{A} \subseteq \mathcal{M}(\mu^*)$. Hence also $\sigma(\mathcal{A}) = \mathcal{F} \subseteq \mathcal{M}(\mu^*)$, and $\mu = \mu^*|_{\mathcal{F}}$ is a measure on (Ω, \mathcal{F}) .

We now turn to the uniqueness of the extension. First suppose that $\mu_0 \Omega < \infty$, and suppose that ν is another extension of μ_0 to \mathcal{F} . Since $\Omega \in \mathcal{A}$, we have $\nu \Omega = \mu \Omega$. We must show that $\mu F = \nu F$ for all $F \in \mathcal{F}$. Let $(A_n)_n$ be a \mathcal{A} -cover for F . then

$$\nu F \leq \sum_n \nu A_n = \sum_n \mu_0 A_n$$

because $\nu|_{\mathcal{A}} = \mu_0$. Since $(A_n)_n$ was an arbitrary \mathcal{A} -cover of F , we must have $\nu F \leq \mu^* F = \mu F$, for all $F \in \mathcal{F}$. Thus also $\nu F^c \leq \mu F^c$, and hence

$$\nu \Omega = \nu F + \nu F^c \leq \mu F + \mu F^c = \mu \Omega = \nu \Omega$$

It is now easy to see that $\nu F = \mu F$, for all $F \in \mathcal{F}$.

Finally, assume that μ_0 is σ -finite on (Ω, \mathcal{A}) . then there exists a sequence $A_n \in \mathcal{A}$ such that $A_n \uparrow \Omega$, and such that $\mu A_n < \infty$ (for all $n \in \mathbb{N}$). As above, we can prove that $\nu(F \cap A_n) = \mu(F \cap A_n)$, for all $n \in \mathbb{N}$, and using Propn. 4.3.4, we see that $\nu F = \mu F$ in this case also.

+

4.6.3 Completion of Measure Spaces*

Suppose that $(\Omega, \mathcal{F}, \mu)$ is a measure space. It may be possible to *extend* the measure μ to a class of sets *larger* than \mathcal{F} , where the measure of the added new sets is determined by μ and \mathcal{F} . For example, suppose that

- (i) $F \in \mathcal{F}$ is such that $\mu F = 0$;
- (ii) $A \subseteq F$

then “clearly” $\mu A = 0$ also. However, if $A \notin \mathcal{F}$, then μA isn’t defined. Yet, μA “ought” to be zero. By adding all those sets whose measure “ought” to be zero, we get a new σ -algebra $\bar{\mathcal{F}}$, called the *completion* of \mathcal{F} w.r.t. μ .

Suppose that $(\Omega, \mathcal{F}, \mu)$ is a measure space, and let μ^* be the outer measure generated by μ (cf. Propn. 4.6.8), i.e.

$$\mu^* E = \inf \left\{ \sum_n \mu F_n : (F_n)_n \text{ a sequence of sets in } \mathcal{F} \text{ with } E \subseteq \bigcup_n F_n \right\}$$

Because we are now dealing with a σ -algebra \mathcal{F} , rather than an algebra \mathcal{A} (as in Propn. 4.6.8) we actually have:

Proposition 4.6.10

$$\mu^* E = \min \{ \mu F : F \in \mathcal{F} \wedge E \subseteq F \}$$

□

Exercise 4.6.11 Prove Propn. 4.6.10.

[Hint: Note first that if $E \subseteq \bigcup_n F_n$, where each $F_n \in \mathcal{F}$, and if $F = \bigcup_n F_n$, then

$$\mu^* E \leq \mu F \leq \sum_n \mu F_n$$

Conclude that

$$\mu^* E = \inf \{ \mu F : E \subseteq F \in \mathcal{F} \}$$

Now choose for each $m \in \mathbb{N}$ a $G_m \in \mathcal{F}$ such that $E \subseteq G_m$ and $\mu G_m \leq \mu^* E + \frac{1}{m}$. (Why can we do this?) Put $G = \bigcap_m G_m$, and show that $\mu G = \mu^* E$.]

□

The following lemma is obvious:

Lemma 4.6.12 *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, and let μ^* be the outer measure generated by μ (via Propn. 4.6.8). Then every μ -null set belongs to $\mathcal{M}(\mu^*)$.*

Definition and Proposition 4.6.13 *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Let*

$$\mathcal{N} := \{ N \subseteq \Omega : \exists F \in \mathcal{F} [\mu F = 0 \wedge N \subseteq F] \}$$

be the set of μ -null sets (cf. Definition 6.3.1). Then

$$\bar{\mathcal{F}} := \{ F \cup N : F \in \mathcal{F}, N \in \mathcal{N} \}$$

is a σ -algebra, called the completion of \mathcal{F} w.r.t. μ .

Moreover, if μ^ is the outer measure generated by μ (via Propn. 4.6.8), then $\bar{\mathcal{F}} \subseteq \mathcal{M}(\mu^*)$, and $\mu^*|_{\bar{\mathcal{F}}}$ is a measure on $(\Omega, \bar{\mathcal{F}})$ which extends μ .*

Finally, if $(\Omega, \mathcal{F}, \mu)$ is a σ -finite measure space, then $\bar{\mathcal{F}} = \mathcal{M}(\mu^)$.*

Proof: We first show that $\bar{\mathcal{F}}$ is a σ -algebra. That $\bar{\mathcal{F}}$ is closed under countable unions follows straightforwardly from the fact that both \mathcal{F} and \mathcal{N} are closed under countable unions. To check that $\bar{\mathcal{F}}$ is closed under complementation, suppose that $F \cup N \in \bar{\mathcal{F}}$, where $F \in \mathcal{F}, N \in \mathcal{N}$. Choose $G \in \mathcal{F}$ such that $\mu G = 0$ and $N \subseteq G$. Then

$$(F \cup N)^c = (F \cup G)^c \cup [G - (F \cup N)]$$

Now $(F \cup G)^c \in \mathcal{F}$, and $G - (F \cup N) \in \mathcal{N}$ (being a subset of G). Hence $(F \cup N)^c \in \bar{\mathcal{F}}$, proving that $\bar{\mathcal{F}}$ is a σ -algebra.

Clearly $\bar{\mathcal{F}} = \sigma(\mathcal{F} \cup \mathcal{N})$. Since $\mathcal{M}(\mu^*)$ is a σ -algebra which includes both \mathcal{F} and \mathcal{N} (by Thm. 4.6.7, Propn. 4.6.8 and Lemma 4.6.12) it follows that $\bar{\mathcal{F}} \subseteq \mathcal{M}(\mu^*)$. Since $\mu^*|_{\mathcal{M}(\mu^*)}$ is a measure (cf. Thm. 4.6.7), so is $\mu^*|_{\bar{\mathcal{F}}}$.

Finally, we show that if $(\Omega, \mathcal{F}, \mu)$ is σ -finite, then also $\mathcal{M}(\mu^*) \subseteq \bar{\mathcal{F}}$. Choose $G_n \in \mathcal{F}$ such that $\Omega = \bigcup_n G_n$, with $\mu G_n < \infty$ for all $n \in \mathbb{N}$, and let $A \in \mathcal{M}(\mu^*)$. It suffices to show that each $A \cap G_n \in \bar{\mathcal{F}}$. Hence we may assume that μ is a *finite* measure. By Propn. 4.6.10, there is $F \in \mathcal{F}$ such that $A \subseteq F_n$ and $\mu^* A = \mu F$. Then

$$\mu^*(A) = \mu F = \mu^* F = \mu^*(F \cap A) + \mu^*(F \cap A^c) = \mu^* A + \mu^*(F - A)$$

Hence $\mu^*(F - A) = 0$ (where we use the fact that μ is finite). It follows, again by Propn. 4.6.10, that $F - A$ is a μ -null set. Thus $A = F - (F - A) \in \bar{\mathcal{F}}$.

□

It is easy to see that $\bar{\mathcal{F}} = \sigma(\mathcal{F} \cup \mathcal{N})$, the smallest σ -algebra which contains each member of \mathcal{F} and also all the null sets.

Exercise 4.6.14 Show that if (S, \mathcal{F}, μ) has completion $(S, \bar{\mathcal{F}}, \mu)$, then

$$\bar{\mathcal{F}} = \{A \subseteq S : A \Delta F \text{ is a } \mu\text{-null set, for some } F \in \mathcal{F}\}$$

[Hint: Let \mathcal{N} be the family of null sets, let $\mathcal{G} = \{A \subseteq S : A \Delta F \in \mathcal{N} \text{ for some } F \in \mathcal{F}\}$, and let $\bar{\mathcal{F}} = \sigma(\mathcal{F} \cup \mathcal{N}) = \{F \cup N : F \in \mathcal{F}, N \in \mathcal{N}\}$. First show that $\mathcal{F}, \mathcal{N} \subseteq \mathcal{G}$, and conclude that $\bar{\mathcal{F}} \subseteq \mathcal{G}$. Next, note that if $A \Delta F \in \mathcal{N}$ for some $F \in \mathcal{F}$, then $A = (F - (F - A)) \cup (A - F)$, where $F \in \mathcal{F}$ and $F - A, A - F \in \mathcal{N}$.

□

4.7 Lebesgue Measure

4.7.1 Lebesgue measure on \mathbb{R}

In this section, we construct the natural notion of *length*, namely Lebesgue measure on \mathbb{R} .

First, we define the *Lebesgue outer measure* $\lambda^* : \mathcal{P}(\mathbb{R}) \rightarrow [0, +\infty]$. (cf. Defn. 4.6.6 for definition of *outer measure*.) Given an interval $I \subseteq \mathbb{R}$, define $|I|$ to be the *length* of the interval. Then, for $A \subseteq \mathbb{R}$, define

$$\lambda^*(A) = \inf \left\{ \sum_{n=1}^{\infty} |I_n| : \text{each } I_n \text{ an interval, } A \subseteq \bigcup_n I_n \right\}$$

Remarks 4.7.1 Note that we also have

$$\lambda^*(A) = \inf \left\{ \sum_{n=1}^{\infty} |I_n| : \text{each } I_n \text{ a finite open interval, } A \subseteq \bigcup_n I_n \right\}$$

For let $\bar{\lambda}A = \inf \left\{ \sum_{n=1}^{\infty} |I_n| : \text{each } I_n \text{ a finite open interval, } A \subseteq \bigcup_n I_n \right\}$. Then clearly $\lambda^*A \leq \bar{\lambda}A$. To prove the reverse inequality, fix $A \subseteq \mathbb{R}$, and choose intervals I_n such that $\sum_n |I_n| < \lambda^*A + \frac{\varepsilon}{2}$. If $|I_n| = +\infty$ for some $n \in \mathbb{N}$, then clearly $\lambda^*A = +\infty$, in which case $\bar{\lambda}A = \lambda^*A$ (i.e. there is nothing to prove). Hence we may assume that each I_n is a finite interval. Now let J_n be an open interval such that $I_n \subseteq J_n$ and $|J_n| \leq |I_n| + \varepsilon 2^{-n-1}$. Then each $\bar{\lambda}A \leq \sum_n |J_n| \leq \sum_n |I_n| + \frac{\varepsilon}{2} < \lambda^*A + \varepsilon$. Letting $\varepsilon \downarrow 0$, we see that $\bar{\lambda}A \leq \lambda^*A$, as required. \square

Proposition 4.7.2 λ^* is an outer measure on \mathbb{R} , and $\lambda^*I = |I|$ for every interval I .

Proof: It is clear that λ^* is a monotone increasing non-negative function with $\lambda^*(\emptyset) = 0$. To prove that λ^* is also countably sub-additive, let $A_1, A_2, \dots \subseteq \mathbb{R}$, and fix an arbitrary $\varepsilon > 0$. By definition of λ^* we may, for each $n \in \mathbb{N}$, choose open intervals $I_{n,1}, I_{n,2}, \dots$ such that

$$A_n \subseteq \bigcup_k I_{n,k} \quad \sum_k |I_{n,k}| \leq \lambda^*A_n + \varepsilon 2^{-n} \quad \text{all } n \in \mathbb{N}$$

Then

$$\bigcup_n A_n \subseteq \bigcup_n \bigcup_k I_{n,k} \quad \lambda^*\left(\bigcup_n A_n\right) \leq \sum_n \sum_k |I_{n,k}| \leq \sum_n \lambda^*A_n + \varepsilon$$

Since ε was arbitrary (> 0), it follows that $\lambda^*\left(\bigcup_n A_n\right) \leq \sum_n \lambda^*A_n$.

It remains to show that $\lambda^*I = |I|$ for every interval $I \subseteq \mathbb{R}$. It is obvious that we always have $\lambda^*I \leq |I|$. To prove the reverse inequality, first assume that I is a compact interval (i.e. $I = [a, b]$, for some $-\infty < a \leq b < +\infty$). Choose finite open intervals $(a_1, b_1), (a_2, b_2), \dots$ such that $[a, b] \subseteq \bigcup_n (a_n, b_n)$ — cf. Remarks 4.7.1. By compactness, there is n such that

$$[a, b] \subseteq \bigcup_{k=1}^n (a_k, b_k)$$

We now check by induction on n that $b - a \leq \sum_{k=1}^n (b_k - a_k)$. This is obvious if $n = 1$. Now suppose that the assertion is true for $n - 1$, and let $(a_1, b_1), \dots, (a_n, b_n)$ be a cover for $[a, b]$. Without loss of generality (relabelling if necessary), we may assume that $b \in (a_n, b_n)$. Then $(a_1, b_1), \dots, (a_{n-1}, b_{n-1})$ is a cover of the interval $[a, a_n]$. By induction hypothesis, we have $a_n - a \leq \sum_{k=1}^{n-1} (b_k - a_k)$, and hence

$$b - a = (b - a_n) + (a_n - a) \leq (b_n - a_n) + \sum_{k=1}^{n-1} (b_k - a_k) = \sum_{k=1}^n (b_k - a_k)$$

as required.

It follows that $b - a \leq \sum_{k=1}^{\infty} (b_k - a_k)$. Since the (a_n, b_n) were an arbitrary open covering of $[a, b]$, it follows that $b - a \leq \lambda^*[a, b]$, i.e. that $|I| \leq \lambda^*I$ for every compact interval I .

It remains to deal with intervals that are non-compact. If I is a bounded interval, then there is a compact interval J such that $J \subseteq I$ and $|I| \leq |J| + \varepsilon$ (where we fix $\varepsilon > 0$). It follows that $|I| \leq \lambda^*J + \varepsilon$. Now since also $\lambda^*J \leq \lambda^*I$, we must have $|I| \leq \lambda^*J + \varepsilon \leq \lambda^*I + \varepsilon$. Letting $\varepsilon \downarrow 0$, we see that $|I| \leq \lambda^*I$ if I is a bounded interval.

Finally, if I is an unbounded interval, then $\lambda^*I = +\infty$: Indeed, if $K > 0$ is arbitrary, there is a compact interval $J \subseteq I$ such that $|J| \geq K$. Then

$$\lambda^*I \geq \lambda^*J \geq K$$

Letting $K \uparrow \infty$, we see that $\lambda^*I = +\infty = |I|$ when I is unbounded.

+

Now that we have constructed an outer measure λ^* , it follows by Thm. 4.6.7 that there is a σ -algebra $\mathcal{L}(\mathbb{R})$ on \mathbb{R} such that $\lambda = \lambda^*|_{\mathcal{L}(\mathbb{R})}$ is a measure on $(\mathbb{R}, \mathcal{L}(\mathbb{R}))$. Indeed, we have $\mathcal{L}(\mathbb{R}) := \mathcal{M}(\lambda^*)$, the family of all λ^* -measurable sets. The σ -algebra $\mathcal{L}(\mathbb{R})$ is called the *σ -algebra of all Lebesgue measurable sets*, or the *Lebesgue algebra*, on \mathbb{R} .

Our next aim is to show that $\mathcal{B}(\mathbb{R}) \subseteq \mathcal{L}(\mathbb{R})$, i.e. that every Borel set is Lebesgue measurable.

Proposition 4.7.3 *Every Borel set is Lebesgue measurable.*

Proof: It suffices to prove that every interval of the form $(-\infty, a]$ is Lebesgue measurable, because the collection of intervals of this form generates $\mathcal{B}(\mathbb{R})$. Let $E \subseteq \mathbb{R}$ be arbitrary, and let $I = (-\infty, a]$. Fix $\varepsilon > 0$, and choose intervals I_1, I_2, \dots such that $\sum_n |I_n| \leq \lambda^*E + \varepsilon$. Note that if J is an arbitrary interval, then so are $J \cap I, J \cap I^c$, and $|J| = |J \cap I| + |J \cap I^c|$.

Now

$$\begin{aligned} \lambda^*E + \varepsilon &\geq \sum_n |I_n| \\ &= \sum_n |I_n \cap I| + \sum_n |I_n \cap I^c| \\ &\geq \lambda^*(E \cap I) + \lambda^*(E \cap I^c) \end{aligned}$$

Letting $\varepsilon \downarrow 0$, we see that $\lambda^*E \geq \lambda^*(E \cap I) + \lambda^*(E \cap I^c)$, for all $E \subseteq \mathbb{R}$.

+

We have almost proved the following theorem:

Theorem 4.7.4 *There exists a unique measure λ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $\lambda I = |I|$ for every interval I .*

Proof: By Propn. 4.7.2, λ^* is an outer measure with $\lambda^*I = |I|$ for every interval, and thus it follows by Thm. 4.6.7 that there is a σ -algebra $\mathcal{L}(\mathbb{R})$ on \mathbb{R} such that $\lambda = \lambda^*|_{\mathcal{L}(\mathbb{R})}$ is a measure on $(\mathbb{R}, \mathcal{L}(\mathbb{R}))$. By Propn. 4.7.3, $\mathcal{B}(\mathbb{R}) \subseteq \mathcal{L}(\mathbb{R})$.

It remains to prove uniqueness: Suppose that μ is another measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with the property that $\mu I = |I|$ for all intervals I . Let $I_n = [-n, n]$. Then $\lambda_n := \lambda|_{I_n}, \nu_n := \mu|_{I_n}$ are *finite* measures on $(I_n, \mathcal{B}(I_n))$. Now $\mathcal{C}_n = \{J : J \text{ an interval, } J \subseteq I_n\}$ is a π -system which generates $\mathcal{B}(I_n)$, and λ_n, ν_n agree on \mathcal{C}_n . Hence, by Propn. 4.6.4, λ_n, μ_n agree on $\mathcal{B}(I_n)$.

Now, if $B \in \mathcal{B}(\mathbb{R})$, then by Propn. 4.3.4,

$$\lambda B = \lim_{n \rightarrow \infty} \lambda_n(B \cap I_n) = \lim_{n \rightarrow \infty} \mu_n(B \cap I_n) = \mu B$$

+

Remarks 4.7.5 It is easy to see that Lebesgue measure can be extended to the *extended* real number system $\mathbb{R} = [-\infty, +\infty]$. The Borel algebra on \mathbb{R} is generated by all sets of the form

$$[-\infty, a) \quad (a, b) \quad (b, \infty] \quad a, b \in \mathbb{R}$$

(which forms a base for the topology on $\bar{\mathbb{R}}$). Thus $\mathcal{B}(\bar{\mathbb{R}})$ is the family of all sets of the form

$$B \quad B \cup \{+\infty\} \quad B \cup \{-\infty\} \quad B \cup \{+\infty, -\infty\}$$

where B is an ordinary Borel set.

□

4.7.2 Lebesgue Measure on \mathbb{R}^d

It is possible to modify the construction of Lebesgue measure on \mathbb{R} to obtain a unique measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, also denoted by λ , which assigns to every rectangle its volume: A rectangle in \mathbb{R}^d is a set

$$R = I_1 \times I_2 \times \cdots \times I_d \quad \text{where } I_1, I_2, \dots, I_d \text{ are intervals in } \mathbb{R}$$

The volume of such a rectangle is defined by $\text{vol}(R) = |I_1| \times |I_2| \times \cdots \times |I_d|$.

We can then define a map $\lambda^* : \mathcal{P}(\mathbb{R}^d) \rightarrow [0, +\infty]$ by

$$\lambda^*(A) = \inf \left\{ \sum_{n=1}^{\infty} \text{vol}(R_n) : \text{each } R_n \text{ a rectangle, } A \subseteq \bigcup_n R_n \right\}$$

and show that λ^* is an outer measure, and that every Borel set is λ^* -measurable.

Instead of performing the construction outlined above, we elect to wait until we have constructed products of measure spaces: Given measure spaces $(\Omega_i, \mathcal{F}_i, \mu_i)$, where $i = 1, \dots, n$, it is possible to construct, in a canonical way, a measure space

$$\left(\prod_{i \leq n} \Omega_i, \bigotimes_{i \leq n} \mathcal{F}_i, \prod_i \mu_i \right)$$

It will turn out that

$$\mathcal{B}(\mathbb{R}^d) = \bigotimes_{i \leq d} \mathcal{B}(\mathbb{R}) \quad \lambda^d = \prod_{i \leq d} \lambda^1$$

where λ^d denotes the d -dimensional Lebesgue measure.

Thus it is possible to construct $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda^d)$ directly from $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$, and a repetition of the construction of Lebesgue measure proves unnecessary.

Chapter 5

Measurable Functions and Random Variables

5.1 Definition of Measurable Function

Let $f : A \rightarrow S$ be a map between two sets. Recall that f induces a set map $f^{-1} : \mathcal{P}(S) \rightarrow \mathcal{P}(A)$ between the power sets — in the opposite direction — by

$$f^{-1}[T] = \{a \in A : f(a) \in T\}$$

Remarks 5.1.1 Here is some motivation for the definition of *measurable function*.

Suppose that X is a *random variable* on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, i.e. a function $X : \Omega \rightarrow \mathbb{R}$ which assigns a number $X(\omega)$ to every outcome $\omega \in \Omega$ — we will make this notion more precise shortly. We would like to be able to discuss the probability that $X = 0$, or that X lies between -1 and 1 , etc. Thus we'd like to know

$$\mathbb{P}(X = 0) := \mathbb{P}(\{\omega \in \Omega : X(\omega) = 0\}) = \mathbb{P}(X^{-1}\{0\})$$

$$\mathbb{P}(-1 \leq X \leq 1) := \mathbb{P}(\{\omega \in \Omega : -1 \leq X(\omega) \leq 1\}) = \mathbb{P}(X^{-1}[-1, 1])$$

However, $\mathbb{P}(F)$ makes sense only if $F \in \mathcal{F}$. Thus, in order to be able to discuss the above probabilities, it is necessary that the sets

$$X^{-1}\{0\} := \{\omega \in \Omega : X(\omega) = 0\} \quad X^{-1}[-1, 1] = \{\omega \in \Omega : X(\omega) \in [-1, 1]\}$$

belong to \mathcal{F} .

More generally, given a Borel set B , we want to be able to discuss the probability that the outcome $X(\omega)$ belongs to B . For

$$\mathbb{P}(X \in B) := \mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\})$$

to make sense, it is necessary that the set

$$X^{-1}B = \{\omega \in \Omega : X(\omega) \in B\}$$

belongs to \mathcal{F} .

Thus: We can only meaningfully discuss the possible values of the random variable X in a probabilistic setting if $X^{-1}B \in \mathcal{F}$ for every $B \in \mathcal{B}(\mathbb{R})$, i.e. it is necessary that

$$X^{-1}\mathcal{B}(\mathbb{R}) \subseteq \mathcal{F}$$

□

We begin with a little set theory:

Proposition 5.1.2 If $T, T_n \subseteq S$, then

- (a) $f^{-1}[T^c] = (f^{-1}[T])^c$
- (b) $f^{-1}[\bigcup_n T_n] = \bigcup_n f^{-1}[T_n]$
- (c) $f^{-1}[\bigcap_n T_n] = \bigcap_n f^{-1}[T_n]$

If $f : A \rightarrow S$, and \mathcal{S} is a family of subsets of S , we denote by $f^{-1}\mathcal{S}$ the family of subsets of A defined by

$$f^{-1}\mathcal{S} = \{f^{-1}[T] : T \in \mathcal{S}\}$$

Proposition 5.1.3 Suppose that $(A, \mathcal{A}), (S, \mathcal{S})$ are measurable spaces, and that $f : A \rightarrow S$ is a map. Then

- (i) $\mathcal{A}' = f^{-1}\mathcal{S}$ is a σ -algebra on A .
- (ii) $\mathcal{S}' = \{T \subseteq S : f^{-1}[T] \in \mathcal{A}\}$ is a σ -algebra on S .

Exercise 5.1.4 Prove Propn. 5.1.2 and 5.1.3. □

Definition 5.1.5 (1.) Let $(A, \mathcal{A}), (S, \mathcal{S})$ be measurable spaces. A map $A \xrightarrow{f} S$ is said to be \mathcal{A}/\mathcal{S} -measurable if and only if $f^{-1}\mathcal{S} \subseteq \mathcal{A}$ (i.e. $f^{-1}[T] \in \mathcal{A}$ for all $T \in \mathcal{S}$).

If the σ -algebras \mathcal{A}, \mathcal{S} are obvious from context, we simply call f a *measurable function*.

(2.) A measurable function X from a probability space (Ω, \mathcal{F}) to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is called a *random variable*.

A measurable function X from a probability space (Ω, \mathcal{F}) to $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is called a *random vector*.

More generally, any measurable function from a probability space to a measurable space is called a *random element*.

(3.) If S is a topological space, then a measurable function $(S, \mathcal{B}(S)) \xrightarrow{f} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is called a *Borel function*.

We are usually interested in the case where $S = \mathbb{R}$ or \mathbb{R}^d .

Thus we have the following *pullback condition* for measurability:

A function is a measurable iff pullbacks of measurable sets are measurable.

Note the similarity with the definition of *continuous function*: A function f between topological spaces X, Y is continuous iff $f^{-1}[V]$ is an open subset of X whenever V is an open subset of Y , i.e. iff pullbacks of open sets are open.

Remarks 5.1.6 (1) The notion of *measure* does not occur in the definition of measurable function/random variable: Only the *measurable spaces* (= set + σ -algebra) play a role.

(2) If $(A, \mathcal{A}) \xrightarrow{f} (S, \mathcal{S})$ is measurable, then

$$A \xrightarrow{f} S \quad \mathcal{S} \xrightarrow{f^{-1}} \mathcal{A}$$

(3) If X is a random variable on (Ω, \mathcal{F}) and $B \subseteq \mathbb{R}$, we write

$$\{X \in B\} \quad \text{for} \quad X^{-1}B = \{\omega \in \Omega : X(\omega) \in B\}$$

(4) We will also allow extended real-valued maps: $A \xrightarrow{f} \bar{\mathbb{R}}$, where $\bar{\mathbb{R}} = [-\infty, +\infty]$.

□

Example 5.1.7 Let (S, \mathcal{S}) be a measurable space. For each $A \subseteq S$ define the **indicator function** $I_A : S \rightarrow \mathbb{R}$ by

$$I_A(s) = \begin{cases} 1 & \text{if } s \in A \\ 0 & \text{otherwise} \end{cases}$$

If A is a measurable set, i.e. $A \in \mathcal{S}$, and if $B \in \mathcal{B}(\mathbb{R})$ is a Borel set, then

$$I_A^{-1}[B] = \begin{cases} \emptyset & \text{if neither } 0, 1 \in B \\ A & \text{if } 1 \in B \text{ and } 0 \notin B \\ A^c & \text{if } 0 \in B \text{ and } 1 \notin B \\ S & \text{if both } 0, 1 \in B \end{cases}$$

It follows that $I_A^{-1}[B] \in \mathcal{S}$ for every $B \in \mathcal{B}(\mathbb{R})$, so that I_A is a measurable function.

Similarly, if I_A is a measurable function, then $I_A^{-1}\{1\} = A \in \mathcal{S}$, since $\{1\}$ is a Borel set. Thus:

A is a measurable set if and only if I_A is a measurable function.

□

The above example is important enough to restate as a Proposition:

Proposition 5.1.8 Suppose that (Ω, \mathcal{F}) is a measurable space, and that $A \subseteq \Omega$. Then the indicator $I_A : \Omega \rightarrow \mathbb{R}$ is a measurable function iff A is a measurable set (i.e. I_A is $\mathcal{F}/\mathcal{B}(\mathbb{R})$ -measurable iff $A \in \mathcal{F}$).

□

Exercise 5.1.9 Suppose that Ω is a set, that $\mathcal{F} := \{\emptyset, \Omega\}$ is the trivial σ -algebra on Ω , and that $\mathcal{G} := \mathcal{P}(\Omega)$ is the powerset-algebra on Ω .

(a) Determine all $\mathcal{F}/\mathcal{B}(\mathbb{R})$ -measurable functions $\Omega \xrightarrow{f} \mathbb{R}$.

(b) Determine all $\mathcal{G}/\mathcal{B}(\mathbb{R})$ -measurable functions $\Omega \xrightarrow{f} \mathbb{R}$.

□

Exercise 5.1.10 Suppose that $(\Omega, \mathcal{F}), (S, \mathcal{S})$ are measurable spaces, and that $f : \Omega \rightarrow S$ is \mathcal{F}/\mathcal{S} -measurable.

(1.) Show that if \mathcal{G} is a σ -algebra on Ω such that $\mathcal{F} \subseteq \mathcal{G}$, then f is also \mathcal{G}/\mathcal{S} -measurable.

(2.) Show that if \mathcal{T} is a σ -algebra on S such that $\mathcal{T} \subseteq \mathcal{S}$, then f is also \mathcal{F}/\mathcal{T} -measurable.

□

To check if a function is measurable, it suffices to check the pullback condition on a generating set:

Proposition 5.1.11 Suppose that $(\Omega, \mathcal{F}), (S, \mathcal{S})$ are measurable spaces, and that $\Omega \xrightarrow{f} S$ is a map. Suppose further that \mathcal{C} is a family of subsets of S such that $\sigma(\mathcal{C}) = \mathcal{S}$. Then f is \mathcal{F}/\mathcal{S} -measurable iff $f^{-1}\mathcal{C} \subseteq \mathcal{F}$.

Proof: (\Rightarrow) is obvious: Clearly $f^{-1}\mathcal{C} \subseteq f^{-1}\mathcal{S}$, and if f is measurable, then $f^{-1}\mathcal{S} \subseteq \mathcal{F}$ by definition of measurability.

(\Leftarrow) : Let $\mathcal{T} = \{T \in \mathcal{S} : f^{-1}[T] \in \mathcal{F}\}$. Then $\mathcal{C} \subseteq \mathcal{T}$, by assumption, and \mathcal{T} is a σ -algebra, by Propn. 5.1.2. (Check this!) Hence $\mathcal{S} = \sigma(\mathcal{C}) \subseteq \mathcal{T} \subseteq \mathcal{S}$ i.e. $\mathcal{T} = \mathcal{S}$.

+

Here are some special cases of Propn. 5.1.11:

Corollary 5.1.12 A function $f : (\Omega, \mathcal{F}) \rightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is measurable iff one of the following conditions holds :

- (a) $\{f \leq c\} \in \mathcal{F}$ for all $c \in \mathbb{R}$. (Recall that $\{f \leq c\} := \{\omega : f(\omega) \leq c\}$.)
- (b) $\{f < c\} \in \mathcal{F}$ for all $c \in \mathbb{R}$.
- (c) $\{f \geq c\} \in \mathcal{F}$ for all $c \in \mathbb{R}$.
- (d) $\{f > c\} \in \mathcal{F}$ for all $c \in \mathbb{R}$.

Proof: (a) In Propn. 5.1.11, take $(S, \mathcal{S}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and \mathcal{C} to be the collection of all intervals of the form $(-\infty, c]$. We already know that these intervals generate the Borel algebra on \mathbb{R} .

(b),(c),(d) are proved similarly.

+

Corollary 5.1.13 If X, S are topological spaces and $X \xrightarrow{f} S$ is continuous, then f is $\mathcal{B}(X)/\mathcal{B}(S)$ -measurable.

Proof: In Propn. 5.1.11, take \mathcal{C} to be the collection of all open subsets of \mathbb{R} .

+

Corollary 5.1.14 Any monotone function from \mathbb{R} to \mathbb{R} is a Borel function, i.e. $\mathcal{B}(\mathbb{R})/\mathcal{B}(\mathbb{R})$ -measurable.

Proof: If $\mathbb{R} \xrightarrow{f} \mathbb{R}$ is monotone, then $\{f < c\}$ is an interval (for all $c \in \mathbb{R}$), and thus in $\mathcal{B}(\mathbb{R})$.

+

5.2 Combinations of Measurable Functions

Measurable functions can be combined in a variety of ways to form new measurable functions:

Proposition 5.2.1 (a) Suppose that $f, g : (\Omega, \mathcal{F}) \rightarrow \bar{\mathbb{R}}$ are measurable functions and that $\alpha \in \mathbb{R}$. Then

$$f + g \quad f^2 \quad \alpha f \quad f \cdot g \quad f/g$$

are measurable functions, where we assume $g \neq 0$ on Ω for the case f/g .

(b) If $f_n : (\Omega, \mathcal{F}) \rightarrow \bar{\mathbb{R}}$ are measurable functions for $n \in \mathbb{N}$, then

$$\sup_n f_n \quad \inf_n f_n \quad \limsup_n f_n \quad \liminf_n f_n$$

are measurable.

(c) If $f_n : (\Omega, \mathcal{F}) \rightarrow \bar{\mathbb{R}}$ are measurable functions for $n \in \mathbb{N}$, and if $f_n \rightarrow f$ pointwise on Ω , then f is measurable.

Proof: (a) Suppose that f, g are measurable. First, we show that $f + g$ is measurable. By Propn. 5.1.11 it suffices to show that

$$\{f + g > c\} \in \mathcal{F} \quad \text{for all } c \in \mathbb{R}$$

Now $f(s) + g(s) > c$ iff $f(s) > c - g(s)$ iff $f(s) > q > c - g(s)$ for some $q \in \mathbb{Q}$. Thus

$$\{f + g > c\} = \bigcup_{q \in \mathbb{Q}} \left(\{f > q\} \cap \{g > c - q\} \right)$$

Now $\{f > q\}, \{g > c - q\} \in \mathcal{F}$ because f, g are measurable. Since \mathbb{Q} is a countable set, $\{f + g > c\} \in \mathcal{F}$ also.

Next, we show that f^2 is measurable. This follows easily from Propn. 5.1.11 using the fact that

$$\{f^2 \leq c\} = \begin{cases} \{-\sqrt{c} \leq f \leq \sqrt{c}\} & \text{if } c \geq 0 \\ \emptyset & \text{else} \end{cases}$$

To see that αf is measurable is easy, e.g. if $\alpha > 0$, then $\{\alpha f < c\} = \{f < \frac{c}{\alpha}\}$.

Next, to see that fg is measurable, use the *polarization identity*

$$fg = \frac{1}{4}[(f + g)^2 - (f - g)^2]$$

Finally, to see that $\frac{f}{g}$ is measurable, it suffices to see that $\frac{1}{g}$ is measurable. But if $c > 0$

$$\{\frac{1}{g} < c\} = \left(\{\frac{1}{c} < g\} \cap \{g > 0\} \right) \cup \left(\{\frac{1}{c} > g\} \cap \{g < 0\} \right)$$

Similar arguments work if $c < 0$ or $c = 0$.

(b) Note that

$$\{\sup_n f_n > c\} = \bigcup_n \{f_n > c\}$$

, that $\inf_n f_n = -\sup_n(-f_n)$, that $\limsup_n f_n = \inf_n \sup_{k \geq n} f_k$ and that $\liminf_n f_n = -\limsup_n(-f_n)$.

(c) Note that if $f_n \rightarrow f$, then $f = \limsup_n f_n = \liminf_n f_n$.

Exercise 5.2.2 If $f, g : (S, \mathcal{S}) \rightarrow \bar{\mathbb{R}}$ are measurable functions, then so are

$$f \vee g = \max\{f, g\} \qquad f \wedge g = \min\{f, g\}$$

because $\max\{f, g\} = \sup\{f, g\}$, etc. In particular, if f is measurable, so are

$$f^+ := f \vee 0 = \max\{f, 0\} \qquad f^- := -(f \wedge 0) = \max\{-f, 0\}$$

f^+, f^- are termed, respectively, the positive and negative parts of f — but note that $f^- \geq 0$! Note that

$$f = f^+ - f^- \qquad |f| = f^+ + f^-$$

In particular: if f is measurable, then $|f|$ is measurable.

□

Exercise 5.2.3 Suppose that $\langle f_n \rangle_n$ is a sequence of measurable functions from a measure space (S, \mathcal{S}) to $\bar{\mathbb{R}}$. Prove that the set $\{s \in S : \lim_n f_n(s) \text{ exists}\}$ is measurable.

□

Measurability, like continuity, is preserved under composition:

Proposition 5.2.4 If $(A, \mathcal{A}) \xrightarrow{f} (S, \mathcal{S})$ and $(S, \mathcal{S}) \xrightarrow{g} (T, \mathcal{T})$ are measurable functions, then $(A, \mathcal{A}) \xrightarrow{g \circ f} (T, \mathcal{T})$ is measurable.

Exercise 5.2.5 Prove Propn. 5.2.4.

□

We have already seen that an indicator function I_A is a measurable function iff A is a measurable set. It follows that from Propn. 5.2.1 that linear combinations of such indicators are measurable as well.

Definition 5.2.6 A measurable function $(\Omega, \mathcal{F}) \xrightarrow{f} \bar{\mathbb{R}}$ is called a *simple* function if $\text{ran} f$ is a finite set.

□

Let (Ω, \mathcal{F}) be a measurable space. Recall that a finite or countably infinite sequence $(F_n)_n$ of members of \mathcal{F} is said to form a *partition* of Ω iff (i) The F_n are mutually disjoint (i.e. $A_n \cap A_m = \emptyset$ when $n \neq m$), and (ii) $\bigcup_n F_n = \Omega$.

Proposition 5.2.7 A measurable function $f : (\Omega, \mathcal{F}) \rightarrow \bar{\mathbb{R}}$ is simple iff it is a linear combination of measurable indicator functions:

$$f = \sum_{i=1}^n c_i I_{F_i}$$

Moreover, the sets $F_i \in \mathcal{F}$ can be chosen to form a partition of Ω .

Proof: It is obvious that a function of the form $f = \sum_{i=1}^n c_i I_{F_i}$ (where each $F_i \in \mathcal{F}$) is simple. f can only take on values which are sums of finitely many of the c_i .

Suppose now that f is simple, i.e. that $\text{ran} f = \{c_1, \dots, c_n\}$ is a finite set. Define $F_i = f^{-1}\{c_i\}$ for $i = 1, \dots, n$. Then the F_i form a partition of Ω , and $f = \sum_{i=1}^n c_i I_{F_i}$.

+

Simple functions play an important part in integration theory. Many important results are proved first for simple functions, and then extended to arbitrary measurable functions by taking limits. The next proposition is therefore *extremely* important:

Proposition 5.2.8 (a) For any non-negative measurable function $(\Omega, \mathcal{F}) \rightarrow \bar{\mathbb{R}}^+$ there exists a sequence of simple measurable functions $f_n, n \in \mathbb{N}$ such that $0 \leq f_n \uparrow f$. Moreover, if f is bounded, we can choose the f_n so that $f_n \rightarrow f$ uniformly.

(b) For any measurable function $(\Omega, \mathcal{F}) \rightarrow \bar{\mathbb{R}}$, there is a sequence of simple measurable functions such that $f_n \rightarrow f$. Moreover, if f is bounded, we can choose the f_n so that $f_n \rightarrow f$ uniformly.

Proof: (a) Define

$$f_n(s) := 2^{-n} [2^n f(s)] \wedge n$$

where $[x]$ is the greatest integer $\leq x$. This elegant definition deciphers as follows:

$$f_n := \sum_{k=0}^{n2^n} \frac{k}{2^n} I_{\{k2^{-n} \leq f < (k+1)2^{-n}\}} \wedge n$$

which means that

$$\begin{aligned} \text{If } f(s) \leq n, \text{ then } f_n(s) &= \frac{k}{2^n} \quad \text{exactly when } \frac{k}{2^n} \leq f(s) < \frac{k+1}{2^n} \\ \text{If } f(s) > n, \text{ then } f_n(s) &= n \end{aligned}$$

Thus f_n is simple and non-negative. Moreover, $0 \leq f(s) - f_n(s) < 2^{-n}$.

Next, we show that $\langle f_n \rangle_n$ is an increasing sequence. If $s \in S$, then there is a unique $m \in \mathbb{N}$ such that $m2^{-(n+1)} \leq f(s) < (m+1)2^{-(n+1)}$, i.e. $f_{n+1} = m2^{-(n+1)}$. If m is even, there is $k \in \mathbb{N}$ such that $m = 2k$, in which case $k2^{-n} \leq f(s) < (k+1)2^{-n}$, i.e. $f_n(s) = k2^{-n} = f_{n+1}(s)$. If m is odd, there is $k \in \mathbb{N}$ such that $2k+1$, in which case $k2^{-n} < (2k+1)2^{-(n+1)} \leq f(s) < (k+1)2^{-n}$, i.e. $f_n(s) = k2^{-n} < (2k+1)2^{-(n+1)} = f_{n+1}(s)$. Thus, whether m is even or odd, $f_n(s) \leq f_{n+1}(s)$.

Next, we show that $f_n(s) \rightarrow f(s)$ for all $s \in S$. If $f(s) = +\infty$, then $f_n(s) = n$ for all $n \in \mathbb{N}$, so certainly $f_n(s) \rightarrow f(s)$. If $f(s) < \infty$, choose N such that $f(s) < N$. If $n \geq N$, then $0 \leq f(s) - f_n(s) \leq 2^{-n}$, and thus $|f(s) - f_n(s)| \leq 2^{-n}$. Thus $f_n(s) \rightarrow f(s)$ in this case also.

Finally, if f is bounded, i.e. $f \leq N$ for some $N \in \mathbb{N}$, then we see that $|f(s) - f_n(s)| \leq 2^{-n}$ for all $n \geq N$ and all $s \in S$, i.e. $f_n \rightarrow f$ uniformly.

(b) Now let f be an arbitrary measurable function to $\bar{\mathbb{R}}$. Then f is the difference of two non-negative measurable functions $f = f^+ - f^-$ (cf. Remarks 5.2.2), and thus, as in (a), there exist non-negative simple functions f_n^+, f_n^- such that $f_n^+ \uparrow f^+, f_n^- \uparrow f^-$. Clearly then also $(f_n^+ - f_n^-) \rightarrow f$. Now note that if f is bounded, so are f^+, f^- . If the f_n^+ and f_n^- converge uniformly, then also $(f_n^+ - f_n^-) \rightarrow f$ uniformly.

+

Exercise 5.2.9 Suppose that (Ω, \mathcal{F}) is a measurable space, and that $\mathcal{A} = \{A_n : n \in \mathbb{N}\}$ is a partition of Ω which generates \mathcal{F} , i.e. $\sigma(\mathcal{A}) = \mathcal{F}$. (Recall that each element of \mathcal{F} is then a union of some of the A_n 's.) We show that the measurable functions are precisely those which are constant on the blocks A_n .

- (a) Show that if $f : \Omega \rightarrow \mathbb{R}$ is $\mathcal{F}/\mathcal{B}(\mathbb{R})$ -measurable, then f is constant on each block A_n (i.e. if $\omega_1, \omega_2 \in A_n$ for some $n \in \mathbb{N}$, then $f(\omega_1) = f(\omega_2)$).
- (b) Show, conversely, that if f is constant on each block, then f is $\mathcal{F}/\mathcal{B}(\mathbb{R})$ -measurable.

□

5.3 Measures and σ -algebras from Measurable Functions

The next proposition shows that measures can be *pushed forward* along measurable functions.

Proposition 5.3.1 Suppose that $(\Omega, \mathcal{F}) \xrightarrow{f} (S, \mathcal{S})$ is measurable, and that μ is a (probability) measure on (Ω, \mathcal{F}) . Define a set function μf^{-1} on \mathcal{S} by

$$(\mu f^{-1})(T) = \mu(f^{-1}[T])$$

Then μf^{-1} is a (probability) measure on (S, \mathcal{S}) .

□

Exercise 5.3.2 Prove Propn. 5.3.1.

□

Remarks 5.3.3 If $(\Omega, \mathcal{F}, \mathbb{P}) \xrightarrow{X} \mathbb{R}$ is a random variable, then $\mathbb{P}X^{-1}$ is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, called the *distribution* or *law* of the random variable X . Note that

$$(\mathbb{P}X^{-1})B = \mathbb{P}(X \in B)$$

□

Exercise 5.3.4 (a) Suppose that $(\Omega, \mathcal{F}, \mathbb{P})$ is the *die space*, i.e. $\Omega = \{1, 2, \dots, 6\}$, $\mathcal{F} = \mathcal{P}(\Omega)$ and $\mathbb{P}(\omega) = \frac{1}{6}$ for all $\omega \in \Omega$. Define $X : \Omega \rightarrow \mathbb{R} : \omega \mapsto \omega^2 - 5$. Show that X is $\mathcal{F}/\mathcal{B}(\mathbb{R})$ -measurable, and determine law of X , i.e. the measure $\mathbb{P}X^{-1}$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

- (b) Suppose that $F : \mathbb{R} \rightarrow \mathbb{R} : x \mapsto x^2$. Show that F is a Borel function, and calculate $\lambda F^{-1}[-1, 3]$ (where λ is Lebesgue measure).

□

A measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is said to be *locally finite* iff $\mu(I) < \infty$ for every compact interval I . The next theorem states that there is a one-to-one correspondence between locally finite measures and increasing right-continuous functions.

Theorem 5.3.5 (a) Suppose that $F : \mathbb{R} \rightarrow \mathbb{R}$ is a right-continuous increasing function with $F(0) = 0$. There is a unique locally finite measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with the property that

$$\mu(a, b] = F(b) - F(a) \quad -\infty < a < b < \infty$$

The measure μ is called the Lebesgue–Stieltjes measure associated with F .

- (b) Conversely, given a locally finite measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, there is a unique right-continuous increasing function F with $F(0) = 0$ so that

$$F(b) - F(a) = \mu(a, b] \quad -\infty < a < b < \infty$$

Exercise 5.3.6 We prove Thm. 5.3.5.

(a) Suppose that F is right-continuous increasing with $F(0) = 0$. Define a function $g : \mathbb{R} \rightarrow \bar{\mathbb{R}}$ by

$$g(t) = \inf\{s \in \mathbb{R} : F(s) \geq t\} \quad t \in \mathbb{R}$$

(Recall that $\inf \emptyset := \infty$.)

(a.1) Show that $g(t) \leq x$ iff $t \leq F(x)$, so that g is a *generalized inverse* of F .

(a.2) Show that g is increasing and left-continuous.

(a.3) Explain why g is a Borel function.

(a.4) Define a measure μ on $(\bar{\mathbb{R}}, \mathcal{B}(\bar{\mathbb{R}}))$ by $\mu := \lambda g^{-1}$, where λ is Lebesgue measure. Use (a.1) to show that $\mu(a, b] = F(b) - F(a)$ whenever $-\infty < a < b < \infty$.

(a.5) Now prove the uniqueness of μ : Explain why if ν is any other measure on \mathbb{R} that satisfies $\nu(a, b] = F(b) - F(a)$ for all $-\infty < a < b < \infty$, then $\nu = \mu$.

(b) Suppose that μ is a locally finite measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Define

$$F(x) := \begin{cases} \mu(0, x] & \text{if } x \geq 0 \\ -\mu(x, 0] & \text{if } x < 0 \end{cases}$$

(b.1) Show that F is right-continuous increasing with $F(0) = 0$.

(b.2) Show that $\mu(a, b] = F(b) - F(a)$ whenever $-\infty < a < b < \infty$.

(b.3) Show that F is the unique function satisfying (b.1) and (b.2).

We can also *pull back* σ -algebras along measurable functions. We already introduced the notion of a σ -algebra generated by a family of sets. We can use this to define the notion of a σ -algebra generated by a random variable.

Definition and Proposition 5.3.7 (a) Let (S, \mathcal{S}) be a measure space, and suppose that \mathcal{X} is a collection of functions $\Omega \rightarrow S$. There is a smallest σ -algebra on Ω denoted by

$$\sigma(\mathcal{X})$$

is the such that all $X \in \mathcal{X}$ are $\sigma(\mathcal{X})/\mathcal{S}$ -measurable. $\sigma(\mathcal{X})$ is called the σ -algebra generated by \mathcal{X} .

We also write $\sigma(X_i : i \in I)$ for the σ -algebra generated by the family $\mathcal{X} = \{X_i : i \in I\}$.

(b) If X is a measurable function, then

$$\sigma(X) = \{X^{-1}[T] : T \in \mathcal{S}\}$$

Proof: (a) Let

$$\mathcal{C} = \{X^{-1}[T] : X \in \mathcal{X}, T \in \mathcal{S}\}$$

Then \mathcal{C} is a family of subsets of Ω , and clearly $\sigma(\mathcal{X}) = \sigma(\mathcal{C})$. (We already know what is meant by $\sigma(\mathcal{C})$, as \mathcal{C} is a family of sets.)

(b) By (a), $\sigma(X)$ is the smallest σ -algebra which includes the family $\mathcal{C} = \{X^{-1}[T] : T \in \mathcal{S}\}$. However, by Propn. 5.1.3, \mathcal{C} is a σ -algebra, and thus $\mathcal{C} = \sigma(X)$.

5.4 Information

In the probabilistic framework, σ -algebras play the role of carriers of information.

Earlier, we saw that if $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, then

- \mathcal{F} is the set containing all those events for which it can be decided *whether or not* they occurred.
- If \mathcal{C} is a family of events, then $\sigma(\mathcal{C})$ is the set containing all those events for which it can be decided *whether or not* they occurred, given that we can decide all the events in \mathcal{C} .

For a random variable X on a probability space Ω , the σ -algebra $\sigma(X)$ can be interpreted in two ways (which are two sides of the same coin):

- $\sigma(X)$ is the *information* carried by X : It is the set of all events that can be decided, given that we know value of X .
- It is the smallest amount of information that we need in order to know the value of X .

Example 5.4.1 For example, consider the experiment of rolling a die, so that $\Omega = \{1, 2, \dots, 6\}$ and $\mathcal{F} = \mathcal{P}(\Omega)$. Let the random variable $X : \Omega \rightarrow \mathbb{R}$ be defined by

$$X(\omega) = \begin{cases} 0 & \text{if } \omega \text{ is even} \\ 1 & \text{if } \omega \text{ is odd} \end{cases}$$

It is easy to check that

$$\sigma(X) = \{\emptyset, \{1, 3, 5\}, \{2, 4, 6\}, \Omega\}$$

Let's consider our interpretations (i) and (ii) above:

- If we know the value of X , all we know is whether the outcome of rolling the die is an even number or an odd number, i.e. all we can decide is whether $\{2, 4, 6\}$ or $\{1, 3, 5\}$ occurred (in addition to being able to decide the certain and impossible events).
- To know the value of X , all we need to know is whether the outcome of the die roll was even or odd. We do not need to know the exact outcome of rolling the die.

□

Example 5.4.2 ADD EXAMPLE ABOUT STOCK PRICE EVOLUTION, FILTRATION, ETC.

□

Exercise 5.4.3 Suppose that $X : (\Omega, \mathcal{F}) \rightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is a function, and that $\omega_1, \omega_2 \in \Omega$ are two elements with the following property:

$$\text{For all } F \in \mathcal{F} \text{ we have } \omega_1 \in F \Leftrightarrow \omega_2 \in F$$

Show that if X is \mathcal{F} -measurable, then $X(\omega_1) = X(\omega_2)$. Thus if \mathcal{F} cannot distinguish between ω_1 and ω_2 , neither can any \mathcal{F} -measurable random variable.

[Hint: Define $x := X(\omega_1)$ and consider $X^{-1}\{x\}$.]

□

If X, Y are random variables such that $\sigma(Y) \subseteq \sigma(X)$, then the information needed to determine the value of Y is a subset of the information required to determine the value of X . Hence, if we know the value of X , we should also know the value of Y . This suggests that Y is a function of X . The following theorem makes this precise.

Theorem 5.4.4 (*Doob–Dynkin Lemma*)

Suppose that $X_i, Y : (\Omega, \mathcal{F} \rightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R})))$ ($i = 1, \dots, n$) are measurable. Then Y is $\sigma(X_1, \dots, X_n)$ –measurable iff there is a Borel function $\mathbb{R}^n \xrightarrow{h} \mathbb{R}$ such that $Y = h(X_1, \dots, X_n)$.

Proof: (\Leftarrow): We first show that the map $X : \Omega \rightarrow \mathbb{R}^n : \omega \mapsto (X_1(\omega), \dots, X_n(\omega))$ is $\sigma(X_1, \dots, X_n)/\mathcal{B}(\mathbb{R}^n)$ –measurable. By Propn. 5.1.11, it suffices to check that $X^{-1} \prod_{i=1}^n (-\infty, c_i] \in \sigma(X_1, \dots, X_n)$ for all $(c_1, \dots, c_n) \in \mathbb{R}^n$, because the family of these lower orthants generates $\mathcal{B}(\mathbb{R}^n)$. But

$$X^{-1} \prod_{i=1}^n (-\infty, c_i] = \bigcap_{i=1}^n X_i^{-1}(-\infty, c_i]$$

so this is obvious. Now $h(X_1, \dots, X_n) = h \circ X$ is a composition of measurable functions, and hence measurable.

(\Rightarrow): First assume that Y is simple, i.e. $Y = \sum_{j=1}^d y_j I_{A_j}$ for some family of mutually disjoint sets A_j (cf. Propn. 5.2.7). Since Y is assumed to be $\sigma(X_1, \dots, X_n)$ –measurable, we see that each $A_j = Y^{-1}\{y_j\}$ belongs to $\sigma(X_1, \dots, X_n)$. Define $X = (X_1, \dots, X_n)$, as above. Reasoning as in Propn. 5.3.7, it is easy to see that

$$A \in \sigma(X_1, \dots, X_n) \quad \text{iff} \quad A = X^{-1}B \quad \text{for some } B \in \mathcal{B}(\mathbb{R}^n)$$

and thus $A_j = X^{-1}B_j$ for some $B_j \in \mathcal{B}(\mathbb{R}^n)$. Now define

$$h = \sum_{j=1}^d y_j I_{B_j}$$

Then $h(X_1, \dots, X_n) = Y$, as required.

Now assume that Y is an arbitrary $\sigma(X_1, \dots, X_n)$ –measurable random variable. Choose a sequence of simple random variables Y_k ($k \in \mathbb{N}$) such that $Y_k \rightarrow Y$ pointwise (cf. Propn. 5.2.8). Hence there exist Borel functions f_k such that $Y_k = f_k(X_1, \dots, X_n)$. Let $M = \{x \in \mathbb{R}^n : \langle f_k(x) \rangle_k \text{ converges}\}$. Then $M \in \mathcal{B}(\mathbb{R}^n)$ (e.g. $M = g^{-1}\{0\}$, where $g = \limsup_k f_k - \liminf_k f_k$). Define $\mathbb{R}^n \xrightarrow{f} \mathbb{R}$ by

$$f(x) = \begin{cases} \lim_k f_k(x) & \text{if } x \in M \\ 0 & \text{else} \end{cases}$$

Then $f = \lim_k f_k I_M$ is a Borel function. Now

$$Y(\omega) = \lim_k Y_k(\omega) = \lim_k f_k(X_1(\omega), \dots, X_n(\omega))$$

which implies two things: (i) $(X_1(\omega), \dots, X_n(\omega)) \in M$, and (ii) $Y = f(X_1, \dots, X_n)$, as required.

Chapter 6

Integration

6.1 Definition and Basic Properties

The aim of this section is to define the integral $\int f \, d\mu$ of a measurable function f w.r.t. a measure μ . Why do we want this? Because

Expectation = Integration

Throughout this subsection, let (S, \mathcal{S}, μ) be a measure space, and let $\mathbf{m}\mathcal{S}$ be the set of all measurable functions from (S, \mathcal{S}) to $\bar{\mathbb{R}}$. We will define a (partial) linear functional, also denoted by μ , or by $\int \cdot \, d\mu$, from $\mathbf{m}\mathcal{S}$ to $\bar{\mathbb{R}}$, i.e.

$$\mu = \int \cdot \, d\mu : \mathbf{m}\mathcal{S} \rightarrow \bar{\mathbb{R}} \qquad f \mapsto \mu f = \int f \, d\mu$$

The quantity $\int f \, d\mu$ need not exist for *every* measurable function f . If it does exist, we say that f is *integrable*.

For the map $\mu : \mathbf{m}\mathcal{S} \rightarrow \bar{\mathbb{R}}$ to be an integral, we would like it to satisfy the following properties:

- I. $\int I_A \, d\mu = \mu A$, i.e. $\mu I_A = \mu A$, for every $A \in \mathcal{S}$.
- II. (Linearity) $\int_A \alpha f + \beta g \, d\mu = \alpha \int f \, d\mu + \beta \int_A g \, d\mu$
- III. (Monotonicity) If $f \leq g$ then $\int f \, d\mu \leq \int g \, d\mu$
- IV. (Continuity) Suppose that $f_n \rightarrow f$. Then $\int f_n \, d\mu \rightarrow \int f \, d\mu$.
[Actually, we won't *quite* get this property, but a weaker one.]

Note that (I.) states that the integral μ is, in some sense, an extension of the measure μ : Every measurable set can be identified with a measurable function (the set A is identified with the indicator function I_A). The integral $\int f \, d\mu = \mu f$ can be thought of as extending the measure μ from sets to *functions*.

The definition of the integral proceeds in three steps:

- (i) Define the integral μ on the set $\mathbf{s}\mathcal{S}^+$ of non-negative simple functions .

(ii) Extend the definition to the set $m\mathcal{S}^+$ of all non-negative measurable functions.

(iii) Finally extend the definition to the set $m\mathcal{S}$ of measurable functions.

If φ is a non-negative simple function, there is only one way to define the integral to be consistent with (I.) and (II.): If $\varphi = \sum_{k=1}^n a_k I_{A_k}$, then define

$$\int \varphi d\mu = \sum_{k=1}^n a_k \mu A_k$$

Some things need checking:

Proposition 6.1.1 (a) The definition of $\mu\varphi$ doesn't depend on the representation of φ as a linear combination of indicators, i.e. if $\varphi = \sum_k a_k I_{A_k} = \sum_j b_j I_{B_j}$, then $\sum_k a_k \mu A_k = \sum_j b_j \mu B_j$.

(b) $\mu I_A = \mu A$.

(c) If $\varphi, \psi \in m\mathcal{S}^+$ and $\alpha, \beta > 0$, then $\mu(\alpha\varphi + \beta\psi) = \alpha \mu\varphi + \beta \mu\psi$.

(d) if $\varphi \leq \psi \in m\mathcal{S}^+$, then $\mu\varphi \leq \mu\psi$.

Proof: We may assume that the a_k are all distinct from each other, and that the b_j are all distinct from each other. Thus $A_k, B_j \in \sigma(\varphi)$, the σ -algebra generated by φ . A little hows that there is a representation $\varphi = \sum_m c_m I_{C_m}$ of φ such that $(C_m)_m$ forms a partition of S , and such that each A_k and B_j is a union of some C_m 's.—just let the C_m 's be the blocks of the partition that generates $\sigma(\varphi)$. In particular, for each k, m , either $A_k \cap C_m = \emptyset$, or $C_m \subseteq A_k$. A similar statement holds for the B_j . Also $c_m = \sum_k \{a_k : C_m \subseteq A_k\}$.

(a) We have

$$\begin{aligned} \sum_k a_k \mu A_k &= \sum_k a_k \sum_m \mu(A_k \cap C_m) = \sum_m \sum_k a_k \mu(A_k \cap C_m) \\ &= \sum_m \sum_k \{a_k \mu C_m : C_m \subseteq A_k\} = \sum_m c_m \mu C_m \end{aligned}$$

(b) is obvious.

(c) Suppose that $\varphi = \sum_k a_k I_{A_k}$, $\psi = \sum_j b_j I_{B_j}$, where $(A_k)_k, (B_j)_j$ are partitions of S . Then $\varphi + \psi = \sum_{k,j} (a_k + b_j) I_{A_k \cap B_j}$ and hence

$$\mu(\varphi + \psi) = \sum_{k,j} (a_k + b_j) \mu(A_k \cap B_j) = \sum_k a_k \mu A_k + \sum_j b_j \mu B_j$$

(d) is obvious.

◄

If f is a non-negative measurable function, then (III.) requires that we must have $\int f d\mu \geq \int \varphi d\mu$ whenever φ is simple, with $f \geq \varphi$. We also know that there is a sequence φ_n of simple non-negative functions such that $\varphi_n \uparrow f$. (IV.) dictates that we should then have $\int \varphi_n d\mu \rightarrow \int f d\mu$, and (III.) that $\lim_n \int \varphi_n d\mu = \sup_n \int \varphi_n d\mu$. The most parsimonious choice, therefore, is to *define*:

Definition 6.1.2

$$\int f \, d\mu = \sup \left\{ \int \varphi \, d\mu : f \geq \varphi \in \mathfrak{m}\mathcal{S}^+ \right\}$$

Note that μf may be equal to $+\infty$.

Exercise 6.1.3 (a) If $\varphi = \sum_k a_k \mu A_k$ is non-negative simple, it is also non-negative measurable, and thus we now have *two* definitions of $\mu\varphi$ namely

$$\mu\varphi = \sum_k a_k \mu A_k \quad \text{and} \quad \mu\varphi = \sup \{ \mu\psi : \varphi \geq \psi \in \mathfrak{s}\mathcal{S}^+ \}$$

Show that these two values of $\mu\varphi$ coincide.

(b) Verify (III.) for non-negative measurable functions, i.e. show that if $f \leq g \in \mathfrak{m}\mathcal{S}^+$, then $\mu f \leq \mu g$.

□

Proving that the integral is still *linear*, i.e. that (II.) holds is much more difficult, and requires a version of (IV.) In fact, a weak version of (IV.) forms the foundation for the whole edifice of integration theory:

Theorem 6.1.4 (Monotone Convergence Theorem)

Suppose that $f_n, f \in \mathfrak{m}\mathcal{S}^+$ such that $f_n \uparrow f$. Then $\mu f_n \uparrow \mu f$.

Proof: It is easy to see that $(\mu f_n)_n$ is an increasing sequence, and that each $\mu f_n \leq \mu f$, so that $\lim_n \mu f_n$ exists (in the extended reals) and $\lim_n \mu f_n \leq \mu f$.

Let $f \geq \varphi \in \mathfrak{s}\mathcal{S}^+$, and suppose $\varphi = \sum_k a_k I_{A_k}$, where the A_k are disjoint, and each $a_k > 0$. For $\varepsilon > 0$, define

$$\varphi_n = \sum_k (1 - \varepsilon) a_k I_{A_k \cap \{f_n \geq (1 - \varepsilon) a_k\}}$$

Then φ_n is a non-negative simple measurable function with $\varphi_n \leq f_n$. Hence

$$\mu f_n \geq \mu \varphi_n = (1 - \varepsilon) \sum_k a_k \mu(A_k \cap \{f_n \geq (1 - \varepsilon) a_k\})$$

Note also that

$$A_k \cap \{f_n \geq (1 - \varepsilon) a_k\} \uparrow A_k$$

for if $s \in A_k$, then $a_k = f(s) = \lim_n f_n(s)$, so that $f_n(s) > (1 - \varepsilon) a_k$ if n is sufficiently large, and thus $s \in A_k \cap \{f_n \geq (1 - \varepsilon) a_k\}$ if n is sufficiently large. By continuity properties of measures,

$$\mu(A_k \cap \{f_n \geq (1 - \varepsilon) a_k\}) \uparrow \mu A_k \quad \text{as } n \rightarrow \infty$$

which in turn yields

$$\mu \varphi_n \uparrow (1 - \varepsilon) \sum_k a_k \mu A_k = (1 - \varepsilon) \mu \varphi \quad \text{as } n \rightarrow \infty$$

Now $\mu f_n \geq \mu \varphi_n$ for each $n \in \mathbb{N}$, and thus

$$\lim_n \mu f_n \geq (1 - \varepsilon) \mu \varphi$$

This is true for any non-negative simple $\varphi \leq f$ and any $\varepsilon > 0$. Taking the supremum over those φ , we see that

$$\lim_n \mu f_n \geq (1 - \varepsilon) \sup \{ \mu \varphi : \varphi \leq f \in \mathfrak{s}\mathcal{S}^+ \} = (1 - \varepsilon) \mu f$$

Letting $\varepsilon \rightarrow 0$, we conclude that $\lim_n \mu f_n \geq \mu f$.

⊢

In Propn. 6.1.1(b), Exercise 6.1.3(b) and Thm 6.1.4, we have seen that (I.), (III.) and a weak version of (IV.) hold. We have also verified (II.) for non-negative simple functions (cf. Propn. 6.1.1(c)). Now we can verify that (II.) holds for non-negative measurable functions:

Proposition 6.1.5 *If $f, g \in m\mathcal{S}^+$ and if $\alpha, \beta \geq 0$, then $\mu(\alpha f + \beta g) = \alpha \mu f + \beta \mu g$.*

Proof: Choose sequences $(\varphi_n)_n, (\psi_m)_m$ of non-negative simple functions such that $\varphi_n \uparrow f$, $\psi_n \uparrow g$. Then each $\alpha\varphi_n + \beta\psi_n$ is non-negative simple, and $(\alpha\varphi_n + \beta\psi_n) \uparrow (\alpha f + \beta g)$. Since (II.) holds for simple functions, and by the Monotone Convergence Theorem, we see that

$$\mu(\alpha f + \beta g) = \lim_n \mu(\alpha\varphi_n + \beta\psi_n) = \alpha \lim_n \mu\varphi_n + \beta \lim_n \mu\psi_n = \alpha \mu f + \beta \mu g$$

⊢

It remains to define the integral for arbitrary measurable functions. Recall that if $f \in m\mathcal{S}$, then $f = f^+ - f^-$, where $f^+ = f \vee 0$, $f^- = -f \wedge 0 = (-f) \vee 0$. Since $f^+, f^- \in m\mathcal{S}^+$, the integrals $\mu f^+, \mu f^-$ have already been defined. If we want to preserve linearity, we therefore *must* define μf by

$$\int f \, d\mu = \int f^+ \, d\mu - \int f^- \, d\mu$$

However, here we face a problem: If both $\mu f^+, \mu f^-$ are equal to $+\infty$, we have $\mu f = \infty - \infty$, an indeterminate form.

Definition and Proposition 6.1.6 *A function $f \in m\mathcal{S}$ is said to be μ -integrable iff $\mu|f| < \infty$. The class of all μ -integrable functions is denoted by $\mathcal{L}^1(S, \mathcal{S}, \mu)$. If $f \in \mathcal{L}^1(S, \mathcal{S}, \mu)$, we define*

$$\mu f = \mu f^+ - \mu f^-$$

Then f is integrable iff $\mu f^+, \mu f^- < \infty$.

Proof: Note that $|f| = f^+ + f^-$, so $\mu|f|$ is finite iff both $\mu f^+, \mu f^-$ are finite.

⊢

Definition 6.1.7 If f is an integrable function and A is a measurable set, we define

$$\int_A f \, d\mu := \int f I_A \, d\mu =: \mu(f; A)$$

to be the integral of f over the set A .

Remarks 6.1.8 Later, we will prove the following important fact: If the Riemann integral $\int_a^b f(x) \, dx$ of a function $\mathbb{R} \xrightarrow{f} \mathbb{R}$ exists, then

$$\int_a^b f(x) \, dx = \int_{[a,b]} f \, d\lambda$$

where λ is Lebesgue measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. If the Riemann integral of a function exists, then so does the Lebesgue integral, and the two integrals coincide. This is obvious if f is a simple function, as you can easily check, but the proof for general f is deferred to a later subsection. Note that the Lebesgue integral may exist even when the Riemann integral does not.

□

Obvious, but often useful, are the following facts:

Proposition 6.1.9 (a) If f is measurable, then f is integrable iff $|f|$ is integrable.
 (b) If f, g are measurable, g is integrable and $|f| \leq g$, then f is integrable as well.
 (c) If f is integrable, then $\mu\{f = \pm\infty\} = 0$, i.e. f is finite μ -a.e.
 (d) If f is integrable, then $|\int f d\mu| \leq \int |f| d\mu$.

Proof: (a) is obvious. (b) follows from the fact that $\mu|f| \leq \mu g < \infty$ (because we have (III.), monotonicity, for non-negative measurable functions).

(c) Let $A = \{s \in S : |f(s)| = \infty\}$. Then $nI_A \leq |f|$ for all $n \in \mathbb{N}$, and hence $n \mu A \leq \mu|f|$. Letting $n \rightarrow \infty$, we see that we must have $\mu|f| = +\infty$ if $\mu A > 0$.

(d) follows because $|\mu f| \leq |\mu f^+| + |\mu f^-| = \mu|f|$.

◊

Remarks 6.1.10 Above, we have defined the integral μf only for $f \in \mathcal{L}^1(S, \mathcal{S}, \mu)$, with the result that $-\infty < \mu f < \infty$, i.e. μf is a finite number.. Before, we defined the integral for arbitrary $g \in m\mathcal{S}^+$, but might then have $\mu g = +\infty$. The restriction to \mathcal{L}^1 is to prevent having to deal with the indeterminate form $\infty - \infty$. However, $\infty - c$ and $c - \infty$ are perfectly fine if $c \neq \infty$. So we sometimes can define the integral of a measurable function f in an *extended* sense: If $\mu f^+ = \infty$, but $\mu f^- = c < \infty$, then we say that $\mu f = \infty$, for example. Nevertheless, such an f is not integrable.

□

Exercise 6.1.11 The decomposition $f = f^+ - f^-$ is but one of many ways that f can be decomposed as a difference of non-negative measurable functions. Show that if $f = g - h$ is a difference of non-negative functions, then $\mu f = \mu g - \mu h$. Thus the definition of the integral of f is independent of the representation of f as a difference of non-negative measurable functions.

[Hint: Apply Propn. 6.1.5 to $f^+ + h = g + f^-$.]

□

Looking at our wish list of properties, i.e. (I.)–(IV.), we see that (I.) holds automatically. (III.) (monotonicity) is easy: If $f \leq g$, then $f^+ \leq g^+$ and $f^- \geq g^-$, so $\mu f^+ \leq \mu g^+$ and $\mu f^- \geq \mu g^-$ (because (III.) holds for non-negative measurable functions, cf. Exercise 6.1.3(b)), and hence $\mu f \leq \mu g$.

We finish this subsection by dealing with (II.) (linearity), and leave (IV.) (continuity) to the next section.

Theorem 6.1.12 If $f, g \in \mathcal{L}^1(S, \mathcal{S}, \mu)$, and $\alpha, \beta \in \mathbb{R}$, then

$$\mu(\alpha f + \beta g) = \alpha \mu f + \beta \mu g$$

Proof: It suffices to prove that $\mu(f + g) = \mu f + \mu g$ and that $\mu(\alpha f) = \alpha \mu f$ (for $f, g \in \mathcal{L}^1$ and $\alpha \in \mathbb{R}$). Now

$$f + g = (f^+ + g^+) - (f^- + g^-)$$

is a representation of $f + g$ as a difference of non-negative measurable functions. By Exercise 6.1.11, it follows that $\mu(f + g) = \mu(f^+ + g^+) - \mu(f^- + g^-)$. Propn. 6.1.5 implies that $\mu(f + g) = \mu f + \mu g$.

Similarly, an application of Propn. 6.1.5 and Exercise 6.1.11 to $\alpha f = \alpha f^+ - \alpha f^-$ (if $\alpha \geq 0$), or $\alpha f = (-\alpha)f^- - (-\alpha)f^+$ (if $\alpha < 0$) yields the conclusion that $\mu(\alpha f) = \alpha \mu f$.

⊥

Exercise 6.1.13 Show that $\mathcal{L}^1(S, \mathcal{S}, \mu)$ is a vector space. Also give an example to show that it may not be closed under multiplication.

□

6.2 Lebesgue's Dominated Convergence Theorem

... the Swiss Army Knife of probability theory...

The following proposition serves as stepping stone in the proof of the Dominated Convergence theorem, but is also very useful in other situations.

Proposition 6.2.1 (a) **FATOU'S LEMMA:** If $f_n \in \mathfrak{m}\mathcal{S}^+$ for $n \in \mathbb{N}$, then

$$\mu(\liminf_n f_n) \leq \liminf_n \mu f_n$$

(b) **REVERSE FATOU LEMMA:** Suppose that $f_n \in \mathfrak{m}\mathcal{S}^+$ for $n \in \mathbb{N}$, and that there exists a $g \in \mathcal{L}^1(S, \mathcal{S}, \mu)$ such that each $f_n \leq g$. Then

$$\limsup_n \mu f_n \leq \mu(\limsup_n f_n)$$

Proof: (a) Let $f = \liminf_n f_n$, and define $g_n = \inf_{m \geq n} f_m$. Then $g_n \uparrow f$, and so the Monotone Convergence Theorem implies that $\mu g_n \uparrow \mu f$. Moreover, $\mu g_n \leq \inf_{m \geq n} \mu f_m$ (by monotonicity, (III.)), and so $\mu f = \lim_n \mu g_n \leq \lim_n \inf_{m \geq n} \mu f_m = \liminf_n \mu f_n$.

⊥

Exercise 6.2.2 Prove the Reverse Fatou Lemma by applying Fatou's Lemma to the sequence $g - f_n$. Why do we require that $g \in \mathcal{L}^1$? Cancellation!

□

Remarks 6.2.3 Under suitable conditions, we see that we have

$$\mu \liminf_n f_n \leq \liminf_n \mu f_n \leq \limsup_n \mu f_n \leq \mu \limsup_n f_n$$

This provides a useful mnemonic: *The terms with the limits on the outside (of the integral) are on the inside (of the string of inequalities).*

The mnemonic *Terms with limits on the inside are on the outside* also works.

□

Theorem 6.2.4 (Dominated Convergence Theorem)

Suppose that f_1, f_2, f_3, \dots is a sequence of measurable functions on (S, \mathcal{S}, μ) such that

$\lim_n f_n(s)$ exists for all $s \in S$;

(i) There is a $g \in \mathcal{L}^1(S, \mathcal{S}, \mu)$ such that $|f_n| \leq g$ for all $n \in \mathbb{N}$.

Then the function $f = \lim_n f_n$ is in $\mathcal{L}^1(S, \mathcal{S}, \mu)$, and

$$\mu f = \lim_n \mu f_n$$

Proof: Since $|f_n| \leq g$, the functions $g \pm f_n$ are non-negative measurable functions, and thus by Fatou's lemma, we see that

$$\mu g + \liminf_n (\pm \mu f_n) = \liminf_n \mu(g \pm f_n) \geq \mu(\liminf_n (g \pm f_n)) = \mu(g \pm f) = \mu g \pm \mu f$$

Subtracting $\mu g < \infty$ from both sides, we see that $\liminf_n \mu f_n \geq \mu f$ and that $\liminf_n (-\mu f_n) \geq -\mu f$, and thus that $\limsup_n \mu f_n \leq \mu f$. Combining, we obtain

$$\mu f \leq \liminf_n \mu f_n \leq \limsup_n \mu f_n \leq \mu f$$

—

Remarks 6.2.5 Note that the DCT states that if a sequence of measurable functions is *dominated* by an integrable function, then limit and integral can be interchanged, i.e.

$$\mu(\lim_n f_n) = \lim_n \mu f_n$$

The integral of the limit is the limit of the integrals.

□

Exercise 6.2.6 (1.) Let $f_n = \frac{1}{n} I_{[0, n]}$ for $n \in \mathbb{N}$.

- (a) Show that $f_n \rightarrow 0$ as $n \rightarrow +\infty$.
- (b) Show that $\int f_n d\lambda = 1$ for all $n \in \mathbb{N}$.
- (c) Why does this not contradict
 - (i) the Monotone Convergence Theorem?
 - (ii) Fatou's Lemma?
 - (iii) the Lebesgue Dominated Convergence Theorem?

(2.) Let $f_n = n I_{(0, \frac{1}{n}]}$ for $n \in \mathbb{N}$.

- (a) Find the function $\lim_{n \rightarrow +\infty} f_n$.
- (b) Show that $\lim_{n \rightarrow +\infty} \int f_n d\lambda \neq \int \lim_{n \rightarrow +\infty} f_n d\lambda$.
- (c) Why does this not contradict the Lebesgue Dominated Convergence Theorem?

□

Remarks 6.2.7 Recall Remarks 6.2.3: The integrals $\int f(x) dx$ and $\int f d\lambda$ coincide whenever the former exists.

An oft-used fact in calculus is that $\frac{d}{dt} \int_a^b f(x, t) dx = \int_a^b \frac{\partial f}{\partial t}(x, t) dx$, provided that $\frac{\partial f}{\partial t}$ is bounded — differentiation under the integral sign. This can be justified via the DCT.

Let $G(t) := \int_a^b f(x, t) \mu(dx)$. We want to show that $G'(t_0) = \int_a^b \frac{\partial f}{\partial t}(x, t_0) \mu(dx)$, under certain commonly satisfied conditions: Suppose that there exists a μ -integrable function $M(x)$ such that $|\frac{\partial f}{\partial t}(x, t)| \leq M(x)$ for all x , and all $t \in (t_0 - \delta, t_0 + \delta)$ (where $\delta > 0$). Let $(h_n)_n$ be a non-zero sequence of reals such that $h_n \rightarrow 0$, and such that each $|h_n| < \delta$. Then

$$G'(t_0) = \lim_n \frac{G(t_0 + h_n) - G(t_0)}{h_n} = \lim_n \left[\int_a^b \frac{f(x, t_0 + h_n) - f(x, t_0)}{h_n} \mu(dx) \right] = \lim_n \int_a^b g_n(x) \mu(dx)$$

where $g_n(x) := \frac{f(x, t_0 + h_n) - f(x, t_0)}{h_n}$. Note that $g_n(x) \rightarrow \frac{\partial f}{\partial t}(x, t_0)$. We claim that the sequence g_n is dominated by M . Indeed, by the Mean Value Theorem, there is, for each x and each $n \in \mathbb{N}$, a $t_x^n \in (t_0 - |h_n|, t_0 + |h_n|) \subseteq (t_0 - \delta, t_0 + \delta)$ such that $g_n(x) = \frac{\partial f}{\partial t}(x, t_x^n)$, and thus $|g_n(x)| \leq M(x)$. Since M is μ -integrable, $\lim_n \int g_n(x) \mu(dx) = \int \lim_n g_n(x) \mu(dx)$, and we are done.

□

6.3 Measure Zero

Suppose that $(\Omega, \mathcal{F}, \mu)$ is a measure space. It may be possible to *extend* the measure μ to a class of sets *larger* than \mathcal{F} , where the measure of the added new sets is determined by μ and \mathcal{F} . For example, suppose that

- (i) $F \in \mathcal{F}$ is such that $\mu F = 0$;
- (ii) $A \subseteq F$

then “clearly” $\mu A = 0$ also. However, if $A \notin \mathcal{F}$, then μA isn’t defined. Yet, μA “ought” to be zero. By adding all those sets whose measure “ought” to be zero, we get a new σ -algebra $\bar{\mathcal{F}}$, called the *completion* of \mathcal{F} w.r.t μ .

Definition 6.3.1 Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, and let $A \subseteq \Omega$.

- (a) We say that A is μ -null if there exists $B \in \mathcal{F}$ such that $A \subseteq B$ and $\mu B = 0$.
(It is not necessary that $A \in \mathcal{F}$.)
- (b) The measure space $(\Omega, \mathcal{F}, \mu)$ is said to be complete iff every μ -null set is measurable, i.e. belongs to \mathcal{F} .

Exercise 6.3.2 Show that a countable union of μ -null sets is μ -null, i.e. that if N_n are μ -null sets, for $n \in \mathbb{N}$, then $\bigcup_n N_n$ is also a μ -null set. □

Definition and Proposition 6.3.3 Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Let

$$\mathcal{N} := \{N \subseteq \Omega : \exists F \in \mathcal{F} [\mu F = 0 \wedge N \subseteq F]\}$$

be the set of μ -null sets (cf. Definition 6.3.1).

- (a) The family of sets

$$\bar{\mathcal{F}} := \{F \cup N : F \in \mathcal{F}, N \in \mathcal{N}\}$$

is a σ -algebra, called the completion of \mathcal{F} w.r.t. μ .

- (b) We have

$$G \in \bar{\mathcal{F}} \text{ iff there are } F_1, F_2 \in \mathcal{F} \text{ such that } F_1 \subseteq G \subseteq F_2 \text{ and } \mu(F_1) = \mu(F_2)$$

- (c) We can extend the measure μ to a measure $\bar{\mu}$ on the σ -algebra $\bar{\mathcal{F}}$ in the obvious way:

$$\text{If } G = F \cup N, \text{ where } F \in \mathcal{F}, N \in \mathcal{N}, \text{ define } \bar{\mu}(G) := \mu(F)$$

- (d) The space $(\Omega, \bar{\mathcal{F}}, \bar{\mu})$ is complete.

Proof: (a) We first show that $\bar{\mathcal{F}}$ is a σ -algebra. That $\bar{\mathcal{F}}$ is closed under countable unions follows straightforwardly from the fact that both \mathcal{F} and \mathcal{N} are closed under countable unions.

To check that $\bar{\mathcal{F}}$ is closed under complementation, suppose that $F \cup N \in \bar{\mathcal{F}}$, where $F \in \mathcal{F}$, $N \in \mathcal{N}$. Choose $G \in \mathcal{F}$ such that $\mu G = 0$ and $N \subseteq G$. Then

$$(F \cup N)^c = (F \cup G)^c \cup [G - (F \cup N)]$$

Now $(F \cup G)^c \in \mathcal{F}$, and $G - (F \cup N) \in \mathcal{N}$ (being a subset of G). Hence $(F \cup N)^c \in \bar{\mathcal{F}}$, proving that $\bar{\mathcal{F}}$ is a σ -algebra. Clearly $\bar{\mathcal{F}} = \sigma(\mathcal{F} \cup \mathcal{N})$.

(b) If $F_1, F_2 \in \mathcal{F}$ are such that $\mu F_1 = \mu F_2$, and if $F_1 \subseteq G \subseteq F_2$, then $G = F_1 \cup (G - F_1)$, where $(G - F_1) \subset (F_2 - F_1)$, so that $G - F_1 \in \mathcal{N}$. It follows that $G \in \bar{\mathcal{F}}$.

Next, if $G \in \bar{\mathcal{F}}$, then (by definition of $\bar{\mathcal{F}}$) there is $F_1 \in \mathcal{F}, N \in \mathcal{N}$ such that $G = F_1 \cup N$. Also, there is $F \in \mathcal{F}$ such that $\mu F = 0$ and $G \subseteq F$. If we now define $F_2 := F_1 \cup F$, we see that $F_1, F_2 \in \mathcal{F}$, with $F_1 \subseteq G \subseteq F_2$, and $\mu F_1 = \mu F_2$. Thus

$$\bar{\mathcal{F}} = \{G \subseteq \Omega : \exists F_1, F_2 \in \mathcal{F} (F_1 \subseteq G \subseteq F_2 \wedge \mu(F_2 - F_1) = 0)\}$$

(c) We need to verify two things: That the extension $\bar{\mu}$ of μ is well-defined on $\bar{\mathcal{F}}$, and that it is a measure. To see that it is well-defined, suppose that $G = F_1 \cup N_1 = F_2 \cup N_2$ are two representations of G , where $F_1, F_2 \in \mathcal{F}$, $N_1, N_2 \in \mathcal{N}$. We must show that $\mu F_1 = \mu F_2$. But $F_1 = F_1 \cap (F_1 \cup N_1) = F_1 \cap (F_2 \cup N_2) = (F_1 \cap F_2) \cup (F_1 \cap N_2)$. It follows easily that $\mu F_1 = \mu(F_1 \cap F_2)$. Similarly $\mu F_2 = \mu(F_1 \cap F_2)$, and hence $\mu F_1 = \mu(F_1 \cap F_2) = \mu F_2$.

Next, we show that $\bar{\mu}$ is a measure on $\bar{\mathcal{F}}$: Suppose that $G_n, (n \in \mathbb{N})$ are mutually disjoint members of $\bar{\mathcal{F}}$. Choose $F_n \in \mathcal{F}, N_n \in \mathcal{N}$ such that $G_n = F_n \cup N_n$ (for $n \in \mathbb{N}$). Then $\bigcup_n G_n = F \cup N$, where $F := \bigcup_n F_n \in \mathcal{F}$ and $N := \bigcup_n N_n \in \mathcal{N}$ (because a countable union of μ -null sets is μ -null). Then by definition of the extension of μ on $\bar{\mathcal{F}}$, we have $\bar{\mu} \bigcup_n G_n = \mu F = \mu \bigcup_n F_n = \sum_n \mu F_n = \sum_n \bar{\mu} G_n$, where we used the fact that μ is a measure on \mathcal{F} to deduce that $\mu \bigcup_n F_n = \sum_n \mu F_n$.

(d) Suppose that N is a null set for $(\Omega, \bar{\mathcal{F}}, \bar{\mu})$. Then there exists $G \in \bar{\mathcal{F}}$ such that $N \subseteq G$ and such that $\bar{\mu}(G) = 0$. There exist therefore a $F \in \mathcal{F}$ $G \subseteq F$ and $\mu F = 0$. Putting all this together, we see that $N \subseteq F$ and $\mu F = 0$, and thus that $N \in \mathcal{N} \subseteq \bar{\mathcal{F}}$. Thus every $\bar{\mu}$ -null set belongs to $\bar{\mathcal{F}}$, as required.

+

Exercise 6.3.4 Show that if (S, \mathcal{F}, μ) has completion $(S, \bar{\mathcal{F}}, \mu)$, then

$$\bar{\mathcal{F}} = \{A \subseteq S : A \Delta F \text{ is a } \mu\text{-null set, for some } F \in \mathcal{F}\}$$

[Hint: Let \mathcal{N} be the family of null sets, let $\mathcal{G} = \{A \subseteq S : A \Delta F \in \mathcal{N} \text{ for some } F \in \mathcal{F}\}$, and let $\bar{\mathcal{F}} = \sigma(\mathcal{F} \cup \mathcal{N}) = \{F \cup N : F \in \mathcal{F}, N \in \mathcal{N}\}$. First show that $\mathcal{F}, \mathcal{N} \subseteq \mathcal{G}$, and conclude that $\bar{\mathcal{F}} \subseteq \mathcal{G}$. Next, note that if $A \Delta F \in \mathcal{N}$ for some $F \in \mathcal{F}$, then $A = (F - (F - A)) \cup (A - F)$, where $F \in \mathcal{F}$ and $F - A, A - F \in \mathcal{N}$.

□

Definition 6.3.5 We shall say that a statement Φ holds *μ -almost everywhere* (or *μ -almost surely* if μ is a probability measure), if the set $\{\omega \in \Omega : \Phi(\omega) \text{ is not true}\}$ where Φ fails to hold is a μ -null set.

First note that completing a measure space does not create any interesting new measurable functions:

Proposition 6.3.6 *Let $(S, \mathcal{S}^\mu, \mu)$ be the completion of (S, \mathcal{S}, μ) . Then a function $S \xrightarrow{f} \bar{\mathbb{R}}$ is \mathcal{S}^μ -measurable iff there is an \mathcal{S} -measurable function $S \xrightarrow{g} \bar{\mathbb{R}}$ such that $f = g$ μ -a.e.*

Proof: (\Rightarrow): First suppose that $f = I_A$ is an indicator function. By Exercise 6.3.4, we know that $\mathcal{S}^\mu = \{A \subseteq S : \exists F \in \mathcal{S} (A \Delta F \text{ is } \mu\text{-null})\}$. So if I_A is \mathcal{S}^μ -measurable, then $I_A = I_F$ μ -a.e. for some $F \in \mathcal{S}$, where then I_F is \mathcal{S} -measurable.

It is now straightforward to see that the proposition holds for simple functions as well.

If f is an arbitrary \mathcal{S}^μ -measurable function, we may choose a sequence f_n of simple \mathcal{S}^μ -measurable functions such that $f_n \rightarrow f$. Then choose simple \mathcal{S} -measurable functions g_n such that $f_n = g_n$ μ -a.e., for all $n \in \mathbb{N}$. Let $g = \limsup_n g_n$. Then $f = g$ μ -a.e. (because $\{s \in S : f(s) \neq g(s)\} \subseteq \bigcup_n \{s \in S : f_n(s) \neq g_n(s)\}$, a countable union of null sets).

(\Leftarrow): Suppose that $f = g$ μ -a.e. for some \mathcal{S} -measurable g . If B is a Borel set, then $f^{-1}(B) \Delta g^{-1}(B) \subseteq \{s \in S : f(s) \neq g(s)\}$ is a μ -null set. Since $g^{-1}(B) \in \mathcal{S}$, we see that $f^{-1}(B) \in \mathcal{S}^\mu$ (by Exercise 6.3.4)

—

Remarks 6.3.7 If f is \mathcal{S} -measurable and if $f = g$ μ -a.e., we cannot generally conclude that g is also \mathcal{S} -measurable. That conclusion is valid, however, if \mathcal{S} is complete w.r.t. μ , i.e. if $\mathcal{S} = \mathcal{S}^\mu$.

□

Next note that two functions which are equal μ -a.e. have the same integrals.

Lemma 6.3.8 *On (S, \mathcal{S}, μ) , if $h \geq 0$ is measurable, then $\mu h = 0$ iff $h = 0$ μ -a.e.*

Proof: The statement is obviously true if h is simple non-negative. For general $h \in m\mathcal{S}^+$, choose simple h_n such that $0 \leq h_n \uparrow h$. If $\mu h = 0$, then by the MCT, $0 \leq \mu h_n \leq \mu h = 0$, so that, by the above, $h_n = 0$ μ -a.e. Thus also $h = \lim_n h_n = 0$ μ -a.e. Conversely, if $h = 0$ μ -a.e., then also $h_n = 0$ μ -a.e., and hence $\mu h = \lim_n \mu h_n = 0$, by the MCT.

—

Theorem 6.3.9 *On (S, \mathcal{S}, μ) , if f, g are measurable functions such that $f = g$ μ -a.e., and if f is integrable (in the extended sense), then g is integrable (in the extended sense), and $\mu f = \mu g$.*

Proof: We have $0 \leq |\mu f - \mu g| \leq \mu |f - g|$, by Propn. 6.1.9. But $f = g$ μ -a.e. iff $|f - g| = 0$ μ -a.e., so Lemma 6.3.8 shows that $0 \leq |\mu f - \mu g| \leq 0$.

—

We can use this to improve the convergence theorems. For example:

Theorem 6.3.10 (Dominated Convergence Theorem)
Suppose that f_1, f_2, f_3, \dots is a sequence of measurable functions on a complete measure space (S, \mathcal{S}, μ) such that

(i) $\lim_n f_n(s)$ exists for μ -a.e. $s \in S$;

(ii) *There is a $g \in \mathcal{L}^1(S, \mathcal{S}, \mu)$ such that $|f_n| \leq g$ μ -a.e. for all $n \in \mathbb{N}$.*

Define f by $f(s) = \lim_n f_n(s)$ if this limit exists, and let $f(s)$ be arbitrary otherwise. Then $f \in \mathcal{L}^1(S, \mathcal{S}, \mu)$, and

$$\mu f = \lim_n \mu f_n$$

Proof: Let

$$N = \{s \in S : \lim_n f_n(s) \text{ does not exist}\} \cup \{s \in S : |f_n(s)| > g(s)\}$$

Then N is a null set, and thus in \mathcal{S} (because the measure space is assumed complete). Define

$$\bar{f}_n = f_n I_{N^c} \quad \bar{g} = g I_{N^c} \quad \bar{f} = f I_{N^c}$$

These functions are also \mathcal{S} -measurable, and we have $\mu \bar{g} = \mu g < \infty$, and

$$\lim_n \bar{f}_n(s) = \bar{f}(s) \quad |\bar{f}_n(s)| \leq \bar{g}(s) \quad \text{for all } s \in S$$

By Theorem 6.2.4, we can conclude that \bar{f} is integrable, and that $\mu \bar{f} = \lim_n \mu \bar{f}_n$. But $\mu f = \mu \bar{f}$ and $\mu f_n = \mu \bar{f}_n$, by Theorem 6.3.9.

—

6.4 Chain Rule, Change of Variables

Here is another way of obtaining new measures from old:

Definition and Proposition 6.4.1 Suppose that $(S, \mathcal{S}, \mu) \xrightarrow{f} (\bar{\mathbb{R}}, \mathcal{B}(\bar{\mathbb{R}}))$ is a non-negative measurable function. Define a set mapping $f \cdot \mu : \mathcal{S} \rightarrow \bar{\mathbb{R}}$ by

$$(f \cdot \mu)A := \int_A f \, d\mu = \mu(f I_A)$$

Then $\nu = f \cdot \mu$ is a measure on (S, \mathcal{S}) .

f is called the μ -density of ν , and also written as the Radon–Nikodým derivative $f = \frac{d\nu}{d\mu}$.

Proof: We need only check that $f \cdot \mu$ is countably additive. Suppose that $A = \bigcup_n A_n$ is a union of a family of mutually disjoint members of \mathcal{S} . Put $f_n = \sum_{k \leq n} f I_{A_k}$. Then $f_n \uparrow f I_A$, and so $\mu f_n \uparrow \mu(f I_A) = (f \cdot \mu)A$, by the MCT. But $\mu f_n = \sum_{k \leq n} \mu(f I_{A_k}) = \sum_{k \leq n} (f \cdot \mu)A_k$, and thus $\mu f_n \uparrow \sum_k (f \cdot \mu)A_k$ (as $n \rightarrow \infty$). We conclude that $(f \cdot \mu)A = \sum_k (f \cdot \mu)A_k$.

—

The following proposition explains the notation $\frac{d\nu}{d\mu}$:

Proposition 6.4.2 (Chain Rule)

On (S, \mathcal{S}, μ) if $S \xrightarrow{f} \bar{\mathbb{R}}^+$ and $S \xrightarrow{g} \bar{\mathbb{R}}$ are measurable, then

$$\mu(fg) = (f \cdot \mu)g$$

i.e. if $\nu = (f \cdot \mu)$ (so that $f = \frac{d\nu}{d\mu}$), then

$$\int fg \, d\mu = \int g \frac{d\nu}{d\mu} \, d\mu = \int g \, d\nu$$

whenever one of these sides exists (in which case the other side exists as well, and the two sides are equal.)

Proof: If $g = I_A$ is an indicator function, then $\mu(fI_A) = (f \cdot \mu)I_A$, by definition of $f \cdot \mu$. If $g = \sum_{k \leq n} \alpha_k I_{A_k}$ is simple, then $\mu(f \sum_{k \leq n} \alpha_k I_{A_k}) = \sum_{k \leq n} \alpha_k \mu(fI_{A_k}) = \sum_{k \leq n} \alpha_k (f \cdot \mu)I_{A_k} = (f \cdot \mu)(\sum_{k \leq n} \alpha_k I_{A_k})$, by linearity of the integral. So the result holds for simple g .

If g is a non-negative measurable function, we may choose simple $g_n \uparrow g$ (cf. Propn. 5.2.8). Then by the MCT, $\mu(fg) = \lim_n \mu(fg_n) = \lim_n (f \cdot \mu)g_n = (f \cdot \mu)g$.

Finally, if g is an arbitrary measurable function, then $\mu|fg| = \mu(f|g|) = (f \cdot \mu)|g|$, since $f, |g|$ are non-negative. Hence $\mu(fg)$ exists iff $(f \cdot \mu)g$ exists (by Propn. 6.1.9). Now split g into its positive and negative parts to see that $\mu(fg) = \mu(fg^+) - \mu(fg^-) = (f \cdot \mu)g^+ - (f \cdot \mu)g^- = (f \cdot \mu)g$.

□

Remarks 6.4.3 The above proof illustrates a useful technique, which David Williams¹ calls the *standard machine*. To prove something holds for all integrals of a certain type:

- First show that it holds for indicator functions;
- Use linearity to show that it holds for simple non-negative functions;
- Then use the MCT to lift the result to non-negative measurable functions;
- And finally split an arbitrary measurable f into its positive and negative parts, and use linearity once again.

□

Recall from Propn. 5.3.1 that if $(S, \mathcal{S}, \mu) \xrightarrow{f} (T, \mathcal{T})$ is measurable, then the map

$$\mu f^{-1} : \mathcal{T} \rightarrow \bar{\mathbb{R}} : B \mapsto \mu f^{-1}[B]$$

defines a measure on (T, \mathcal{T}) . Also if $(T, \mathcal{T}) \xrightarrow{g} (\bar{\mathbb{R}}, \mathcal{B}(\bar{\mathbb{R}}))$ is measurable, then so is $(S, \mathcal{S}) \xrightarrow{g \circ f} (\bar{\mathbb{R}}, \mathcal{B}(\bar{\mathbb{R}}))$

The next propn. shows that the integrals $\int g \circ f d\mu$ and $\int g d(\mu f^{-1})$ are equal:

Proposition 6.4.4 (Change of Variables)

Given a measure space (S, \mathcal{S}, μ) , a measurable space (T, \mathcal{T}) and two measurable maps $f : S \rightarrow T$ and $g : T \rightarrow \bar{\mathbb{R}}$, then

$$\mu(g \circ f) = (\mu f^{-1})g \quad \text{i.e.} \quad \int g \circ f d\mu = \int g d(\mu f^{-1})$$

whenever one of these sides exists (in which case the other side exists as well, and the two sides are equal.)

□

Exercise 6.4.5 Prove Propn. 6.4.4.

[Hint: Use the standard machine. For arbitrary measurable g , observe that $\mu|g \circ f| = \mu(|g| \circ f) = (\mu f^{-1})|g|$, because $|g|$ is non-negative. This proves that $g \circ f$ is μ -integrable iff g is μf^{-1} -integrable (i.e. one side exists iff the other exists.)]

□

¹cf. his excellent (and short) book *Probability with Martingales*.

6.5 Riemann Integral vs. Lebesgue Integral

Let f be a real-valued function defined and bounded on an interval $[a, b]$. We recall here the definition of the Riemann integral

$$\int_a^b f(t) dt$$

Define a function $I : [a, b] \rightarrow \mathbb{R}$ by

$$I(t) = \int_a^t f(s) ds$$

Let

$$P = \{a = t_0 < t_1 < t_2 < \cdots < t_n = b\}$$

be a partition of $[a, b]$. Choose $t_k^* \in [t_{k-1}, t_k]$. Then we ought to have

$$\begin{aligned} \int_a^b f dG &= I(b) = \sum_{k=1}^n (I(t_k) - I(t_{k-1})) \\ &= \sum_{k=1}^n f(t_k^*)(t_k - t_{k-1}) \end{aligned}$$

The sum in the last line of this equation is called a Riemann sum. This approximation ought to hold (for a sufficiently nice integrand f) provided that the partition P is sufficiently fine, and the smaller the sizes of the $\Delta_k t = t_k - t_{k-1}$, the better the approximation. Let

$$\|P\| = \max\{\Delta_k t : k = 1, \dots, n\}$$

and let $n(P) = n$ (i.e. number of points +1 in P). Define *upper* and *lower Riemann sums* as follows:

$$\begin{aligned} U(f, P) &= \sum_{k=1}^n \sup\{f(t) : t \in [t_k, t_{k-1}]\} \cdot [t_k - t_{k-1}] \\ L(f, P) &= \sum_{k=1}^n \inf\{f(t) : t \in [t_k, t_{k-1}]\} \cdot [t_k - t_{k-1}] \end{aligned}$$

It is obvious that we always have $L(f, P) \leq U(f, P)$.

Next define the *upper* and *lower Riemann integrals* by:

$$\begin{aligned} \overline{\int_a^b} f dt &:= \inf\{U(f, P) : P \text{ is a partition of } [a, b]\} \\ \underline{\int_a^b} f dt &:= \sup\{L(f, P) : P \text{ is a partition of } [a, b]\} \end{aligned}$$

It is easily seen that

$$\underline{\int_a^b} f dt \leq \overline{\int_a^b} f dt$$

A function f is said to be *Riemann integrable* over $[a, b]$ provided that the upper and lower integrals are equal. In that case the Riemann integral is defined to be their common value:

$$\int_a^b f dG := \underline{\int_a^b} f dG = \overline{\int_a^b} f dG$$

Theorem 6.5.1 *Let f be a bounded real-valued function on the compact interval $[a, b]$. Then*

(a) *f is Riemann integrable iff f is continuous λ -a.e.*

(b) *If f is Riemann integrable, then f is Lebesgue integrable, and the integrals are equal:*

$$\int_a^b f \, dt = \int_{[a,b]} f \, d\lambda.$$

Proof: Assume that f is Riemann integrable. Then we can choose a sequence P_n of successively finer partitions of $[a, b]$ such that $U(f, P_n) - L(f, P_n) < \frac{1}{n}$. Define functions g_n, h_n on $[a, b]$ as follows: For each n , $g_n(a) = h_n(a) = f(a)$. If $P_n = \{a = t_0^n < t_1^n < t_2^n < \cdots < t_{m_n}^n = b\}$, then g_n, h_n are step functions, with steps determined by P_n , defined as follows: If $t \in [a, b]$, then $t \in (t_{k-1}^n, t_k^n]$ for some k , and we define

$$g_n(t) = \inf\{f(x) : t_{k-1}^n < x \leq t_k^n\} \quad h_n(t) = \sup\{f(x) : t_{k-1}^n < x \leq t_k^n\}$$

Then g_n, h_n are clearly simple Borel functions, designed so that

$$\int_{[a,b]} g_n \, d\lambda = L(f, P_n) \quad \int_{[a,b]} h_n \, d\lambda = U(f, P_n)$$

Moreover $(g_n)_n$ is a bounded increasing sequence, with $g_n \leq f$, and $(h_n)_n$ is a bounded decreasing sequence, with $h_n \geq f$. Define $g = \lim_n g_n, h = \lim_n h_n$. Then g, h are Borel functions, and by the DCT we have $\int_{[a,b]} g \, d\lambda = \lim_n L(f, P_n) = \int_a^b f \, dt$ and $\int_{[a,b]} h \, d\lambda = \lim_n U(f, P_n) = \int_a^b f \, dt$. Hence $\int_{[a,b]} h - g \, d\lambda = 0$.

Now since $h \geq g$, Lemma 6.3.8 implies that $h = g$ λ -a.e. on $[a, b]$. Since $g \leq f \leq h$, we must have $g = f = h$ λ -a.e., and thus $\int f \, d\lambda = \int g \, d\lambda = \lim_n L(f, P_n) = \int_a^b f \, dt$. This proves (b).

Next, note that if $t \notin \bigcup_n P_n$, and if $h(t) = g(t)$, then f is necessarily continuous at t : For then $g(t) = f(t) = h(t)$, i.e.

$$\liminf_n \{f(x) : t_{k_n-1}^n < x \leq t_{k_n}^n\} = f(t) = \sup_n \inf\{f(x) : t_{k_n-1}^n < x \leq t_{k_n}^n\}$$

(where k_n is such that $t \in (t_{k_n-1}^n, t_{k_n}^n]$) and thus all values of $f(x)$ must lie close to $f(t)$ if x is close to t . Hence any discontinuity of f must belong to $\bigcup_n P_n \cup \{t : g(t) \neq h(t)\}$, a set of λ -measure zero. This shows that if f is Riemann integrable, the f is continuous λ -a.e., proving one direction of (a).

Conversely, suppose that f is continuous λ -a.e. Let P_n be a partition of $[a, b]$ that divides it into 2^n subintervals of equal length, and construct simple Borel functions g_n, h_n as above. If f is continuous at t , then obviously $\lim_n g_n(t) = f(t) = \lim_n h_n(t)$. Hence $\lim_n (h_n - g_n) = 0$ λ -a.e. By the DCT, we see that $0 = \lim_n \int_{[a,b]} h_n - g_n \, d\lambda = \lim_n (U(f, P_n) - L(f, P_n))$, from which Riemann integrability easily follows.

—

Chapter 7

Differentiation

7.1 Bounded Linear Operators

As preliminary to the definition of the derivative of a general multivariate vector-valued function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ we discuss continuity of linear operators. We start with a simple criterion for continuity: A linear operator $L : V \rightarrow W$ between normed vector spaces is said to be *bounded* if and only if there is a constant C such that $\|L(x)\| \leq C\|x\|$ for all $x \in V$. The next proposition shows that a linear operator is continuous if and only if it is bounded:

Proposition 7.1.1 *Let $L : V \rightarrow W$ be a linear operator between normed vector spaces V, W . Then L is continuous if and only if it is a bounded operator, i.e. iff there exists a constant C such that*

$$\|L(x)\| \leq C\|x\| \quad \text{for all } x \in V$$

Proof: Suppose L is continuous. Choose δ so that $\|L(x)\| \leq 1$ whenever $\|x\| \leq \delta$. Let $C \geq \frac{1}{\delta}$. If $x \in V$, then $\|\frac{x}{C\|x\|}\| \leq \delta$, so $\|L(\frac{x}{C\|x\|})\| \leq 1$, i.e. $\|L(x)\| \leq C\|x\|$.

Conversely, suppose that L is a bounded operator, and that $\|L(x)\| \leq C\|x\|$ for all $x \in V$. To show that L is continuous, it suffices to show that $L(x_n) \rightarrow L(x)$ whenever $x_n \rightarrow x$, i.e. that $\|L(x_n) - L(x)\| \rightarrow 0$ whenever $\|x_n - x\| \rightarrow 0$. But this is easy:

$$\|L(x_n) - L(x)\| = \|L(x_n - x)\| \leq C\|x_n - x\| \rightarrow 0 \text{ as } \|x_n - x\| \rightarrow 0$$

◄

Before we prove that linear operators between finite-dimensional vector spaces are continuous, we need a simple Lemma:

Lemma 7.1.2 *If $L : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is a linear transformation, then L is bounded, i.e. there exists a constant C such that*

$$\|L(x)\| \leq C\|x\| \quad \text{for all } x \in \mathbb{R}^n$$

(where the norms are the standard Euclidean norms.)

Proof: Let $\mathbf{e}_1, \dots, \mathbf{e}_n$ denote the standard basis of \mathbb{R}^n , and put $C = n \max\{\|L(\mathbf{e}_1)\|, \dots, \|L(\mathbf{e}_n)\|\}$, so that each $\|L(\mathbf{e}_i)\| \leq \frac{C}{n}$. If $h = (h_1, \dots, h_n)^{tr} \in \mathbb{R}^n$, then $h = \sum_{i=1}^n h_i \mathbf{e}_i$, and each

$|h_i| = \sqrt{h_i^2} \leq \sqrt{h_1^2 + \cdots + h_n^2} = \|h\|$. It follows, using the triangle inequality and the inequalities just obtained that

$$\|L(h)\| = \|L(\sum_{i=1}^n h_i \mathbf{e}_i)\| \leq \sum_{i=1}^n |h_i| \|L(\mathbf{e}_i)\| \leq \sum_{i=1}^n \|h\| \frac{C}{n} = C \|h\|$$

□

Proposition 7.1.1 and Lemma 7.1.2 immediately imply that:

Corollary 7.1.3 Any linear operator $L : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is continuous.

Remarks 7.1.4 1. Suppose that V, W are normed vector spaces, and let $\mathcal{L}(V, W)$ be the set of all bounded linear operators from V to W . It is clearly possible to add linear operators, and to multiply them by scalars:

$$(L_1 + L_2)(x) := L_1(x) + L_2(x) \quad (\lambda L)(x) := \lambda L(x)$$

Further more, the sum of two bounded linear operators is bounded: If $\|L_i(x)\| \leq C_i \|x\|$ for $i = 1, 2$, then $\|(L_1 + L_2)(x)\| \leq (C_1 + C_2)\|x\|$ by the triangle inequality. Similarly, λL is bounded if L is. It follows that $\mathcal{L}(V, W)$ is a vector space. The bounds C can be used to define a norm on $\mathcal{L}(V, W)$ — the *operator norm*:

$$\|L\| = \inf\{C : \|L(x)\| \leq C\|x\| \text{ for all } x\}$$

This plays an important role in functional analysis (but not in this course).

2. As every finite dimensional real vector space is isomorphic to an \mathbb{R}^n , the preceding corollary proves that all linear operators between finite dimensional normed vector spaces are continuous. This breaks down for infinite dimensional vector spaces. Consider, for example, the space $V = \{f : [0, 1] \rightarrow \mathbb{R} : f \text{ is continuous and differentiable on } (0, 1)\}$. This is a normed space, with norm defined by $\|f\| := \max_{x \in [0, 1]} |f(x)|$. For $x_0 \in (0, 1)$ map $D_{x_0} : V \rightarrow \mathbb{R} : f \mapsto f'(x_0)$ is clearly linear, but it is not continuous: Define $f_n(x) := \frac{1}{n} \sin 2\pi n x$. Then $f_n \rightarrow 0$ (as $\|f_n\| \leq \frac{1}{n} \rightarrow 0$). Nevertheless, with $x_0 = \frac{1}{2}$, we have $f'_n(x_0) = 2\pi \cos \pi n$, so $f'_n(x_0) \not\rightarrow 0$, i.e. $D_{x_0} f_n \not\rightarrow D_{x_0} 0$.

□

7.2 The Derivative

7.2.1 Definition of the Derivative

Recall that a function $f : \mathbb{R} \rightarrow \mathbb{R}$ is differentiable at $x_0 \in \mathbb{R}$ if and only if there is a number a such that

$$\lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0} = a, \quad \text{equivalently} \quad \lim_{h \rightarrow 0} \frac{f(x_0 + h) - f(x_0)}{h} = a \quad (*)$$

The number a is called the *derivative of f at x_0* , and denoted $f'(x_0)$.

If we try to extend this definition blindly to functions $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ we run into trouble: $x_0, h, f(x_0)$, etc. are vectors, and one cannot *divide* by vectors. A careful analysis the meaning of $(*)$ is therefore necessary.

Examples 7.2.1 1. Define

$$\varepsilon_{x_0}(x) := \frac{f(x) - f(x_0)}{x - x_0} - a$$

so that

$$f(x) = f(x_0) + a(x - x_0) + \varepsilon_{x_0}(x)(x - x_0)$$

(*) says that $\varepsilon_{x_0}(x) \rightarrow 0$ as $x \rightarrow x_0$. Note that $\varepsilon_{x_0}(x)(x - x_0) \rightarrow 0$ “doubly fast” as $x \rightarrow x_0$: First, because $\varepsilon_{x_0}(x) \rightarrow 0$ as $x \rightarrow x_0$, and secondly because in addition $x - x_0 \rightarrow 0$. Thus for x close to x_0 , we have

$$f(x) = f(x_0) + a(x - x_0) + \text{something very small}$$

In beginners’ courses on calculus, the process of finding a derivative is usually introduced as the process of finding a tangent. Assuming f is “smooth” at x_0 , the tangent is the straight line $y = f(x_0) + a(x - x_0)$ which best approximates the function f in a neighbourhood of x_0 . Thus $\varepsilon_{x_0}(x)(x - x_0)$ is the amount by which the function f deviates from the tangent, and it is “doubly small” for x close to x_0 .

This idea of *linearization* is at the heart of differential calculus. Note that every linear function $L : \mathbb{R} \rightarrow \mathbb{R}$ is given by multiplication by a constant, i.e. $L(x) = ax$ for some $a \in \mathbb{R}$. Indeed, if we define $a := L(1)$, then by linearity, we have $L(x) = L(x \cdot 1) = x \cdot L(1) = ax$. We now see that we have $f(x) = f(x_0) + L(x - x_0) + \varepsilon_{x_0}(x)(x - x_0)$ which means that the change in f at x_0 is roughly linear:

$$f(x) - f(x_0) = L(x - x_0) + \text{something very small}$$

for some linear operator $L : \mathbb{R} \rightarrow \mathbb{R}$.

2. Let’s see if we can extend this idea: Consider a smooth surface in \mathbb{R}^3 , given by a function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$, written as $z = f(\mathbf{x})$, where $\mathbf{x} = (x, y) \in \mathbb{R}^2$. The best “straight” approximation of f near a point $\mathbf{x}_0 \in \mathbb{R}^2$ is given by the tangent plane at that point:

$$z = f(x_0, y_0) + a(x - x_0) + b(y - y_0)$$

(i.e. the tangent plane is $z = ax + by + c$, where $c = f(x_0, y_0) - ax_0 - by_0$.) Thus

$$f(x, y) = f(x_0, y_0) + a(x - x_0) + b(y - y_0) + \text{something very small}$$

for (x, y) near (x_0, y_0) . Here, again, we have *linearization*: Every linear function $L : \mathbb{R}^2 \rightarrow \mathbb{R}$ is given by $L(x, y) = ax + by$ for some constants a, b . Indeed, define $a := L(1, 0)$ and $b := L(0, 1)$. Linearity of L then implies that

$$L(x, y) = L(x(1, 0) + y(0, 1)) = xL(1, 0) + yL(0, 1) = ax + by$$

We thus have

$$f(\mathbf{x}) = f(\mathbf{x}_0) + L(\mathbf{x} - \mathbf{x}_0) + \text{something very small}$$

for some linear operator $L : \mathbb{R}^2 \rightarrow \mathbb{R}$.

□

We are now almost ready to define the notion of derivative for a map $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$. We want to define the derivative $Df(\mathbf{x}_0)$ of f at the point x_0 to be a *linear operator* with the property that

$$f(\mathbf{x}) = f(\mathbf{x}_0) + Df(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \text{something very small}$$

The problem is defining the “something small”. Example 7.2.1(1) points the way:

Definition 7.2.2 Suppose that $U \subseteq \mathbb{R}^n$ and that x_0 is an interior point of U . We say that $f : U \rightarrow \mathbb{R}^m$ is differentiable at x_0 if there is a linear operator $L : \mathbb{R}^n \rightarrow \mathbb{R}^m$ such that

$$f(x) = f(x_0) + L(x - x_0) + \varepsilon_{x_0}(x) \cdot \|x - x_0\| \quad (**)$$

where $\varepsilon : U \rightarrow \mathbb{R}^m$ has the property that $\varepsilon_{x_0}(x) \rightarrow 0$ as $x \rightarrow x_0$.

The linear operator L is called the *derivative* (or *Fréchet derivative*) of f at x_0 , and denoted by $L = Df(x_0)$.

If f is differentiable at every point of U , we say that f is differentiable on U .

Note that $f(x), f(x_0), L(x - x_0), \varepsilon_{x_0}(x) \in \mathbb{R}^m$, whereas $x - x_0 \in \mathbb{R}^n$. It is not possible to multiply vectors $\varepsilon_{x_0}(x)$ and $x - x_0$. We can however, multiply the $\varepsilon_{x_0}(x)$ with the scalar $\|x - x_0\|$. We then maintain the idea that $\varepsilon_{x_0}(x) \cdot \|x - x_0\| \rightarrow 0$ “doubly fast” as $x \rightarrow x_0$: Firstly, because $\varepsilon_{x_0}(x) \rightarrow 0$ as $x \rightarrow x_0$, and secondly because then in addition $\|x - x_0\| \rightarrow 0$ as well.

There’s a loose end we need to tie up immediately: In Definition 7.2.2, we define $Df(x_0)$ to be “the” linear operator satisfying (**). But what if there is more than one such operator? There isn’t, but to prove it, we first need a lemma:

Lemma 7.2.3 Suppose that $L : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is a linear operator, and define $\varepsilon : \mathbb{R}^n \rightarrow \mathbb{R}^m$ by $\varepsilon(h) = \frac{L(h)}{\|h\|}$. If $\varepsilon(h) \rightarrow 0$ as $h \rightarrow 0$, then $L = 0$, the constant map with value 0.

Proof: Note that for $\alpha > 0$ we have

$$L(h) = \alpha^{-1}L(\alpha h) = \alpha^{-1}\varepsilon(\alpha h) \|\alpha h\| = \varepsilon(\alpha h) \|h\|$$

Now $\varepsilon(\alpha h) \rightarrow 0$ as $\alpha \rightarrow 0$, and so $L(h) = 0$.

—

Proposition 7.2.4 The derivative, if it exists is unique, i.e.: Suppose that L_1, L_2 are linear operators satisfying

$$f(x) = f(x_0) + L_i(x - x_0) + \varepsilon_i(x) \cdot \|x - x_0\| \quad i = 1, 2$$

where $\varepsilon_i(x) \rightarrow 0$ as $x \rightarrow x_0$. Then $L_1 = L_2$.

Proof: Subtracting, we see that $L_1(x - x_0) - L_2(x - x_0) + \varepsilon_1(x) \cdot \|x - x_0\| - \varepsilon_2(x) \cdot \|x - x_0\| = 0$, i.e. that

$$L(h) = \varepsilon(h) \|h\|$$

where $L := L_1 - L_2$, $h := x - x_0$, and $\varepsilon(h) := \varepsilon_2(x_0 + h) - \varepsilon_1(x_0 + h)$. Note that $\|\varepsilon(h)\| \leq \|\varepsilon_1(x_0 + h)\| + \|\varepsilon_2(x_0 + h)\|$, and that $\varepsilon_i(x_0 + h) \rightarrow 0$ as $h \rightarrow 0$, so that also $\varepsilon(h) \rightarrow 0$ as $h \rightarrow 0$. By the preceding lemma, $L = 0$ identically, i.e. $L_1 = L_2$.

—

It is often convenient to recast Definition 7.2.2 in the following equivalent form:

Proposition 7.2.5 $f : U \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^m$ is differentiable at x_0 if and only if there is a linear operator $L : \mathbb{R}^n \rightarrow \mathbb{R}^m$ such that

$$\lim_{h \rightarrow 0} \frac{\|f(x_0 + h) - f(x_0) - L(h)\|}{\|h\|} = 0$$

Then $L = Df(x_0)$.

Remark 7.2.6 Note that we may have two different norms in the expression $\frac{\|f(x_0 + h) - f(x_0) - L(h)\|}{\|h\|}$: When $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$, the norm in the numerator is the \mathbb{R}^m -norm, whereas the norm in the denominator is the \mathbb{R}^n -norm.

□

Proof: By definition, f is differentiable at x_0 iff and only if there exist a linear operator L and a map ε such that $f(x) = f(x_0) + L(x - x_0) + \varepsilon(x) \|x - x_0\|$, where $\varepsilon(x) \rightarrow 0$ as $x \rightarrow x_0$. Put $h := x - x_0$. Then we have

$$f(x_0 + h) = f(x_0) + L(h) + \varepsilon(x_0 + h) \|h\|$$

and so

$$\frac{\|f(x_0 + h) - f(x_0) - L(h)\|}{\|h\|} = \|\varepsilon(x_0 + h)\|$$

Taking $\lim_{h \rightarrow 0}$ on both sides yields the result.

+

Once more:

A function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is differentiable at a point $x_0 \in \mathbb{R}^n$ if and only if there exists a linear transformation $L : \mathbb{R}^n \rightarrow \mathbb{R}^m$ with the property that the function

$$\varepsilon : \mathbb{R}^n \rightarrow \mathbb{R}^m : h \mapsto \frac{1}{\|h\|} [f(x_0 + h) - f(x_0) - L(h)]$$

is such that $\varepsilon(h) \rightarrow 0$ as $h \rightarrow 0$.

Then $L = Df(x_0)$.

Example 7.2.7 Consider the function $f : \mathbb{R}^2 \rightarrow \mathbb{R} : (x, y) \mapsto x^2 + y$ at the point $\mathbf{x}_0 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$. With $\mathbf{h} = \begin{pmatrix} h \\ k \end{pmatrix}$ we have

$$f(\mathbf{x}_0 + \mathbf{h}) - f(\mathbf{x}_0) = 4h + k + h^2$$

Now $4h + k$ is linear in h, k and $h^2 \rightarrow 0$ “doubly fast” as $h, k \rightarrow 0$. Define, therefore

$$L(\mathbf{h}) = 4h + k = \begin{pmatrix} 4 & 1 \end{pmatrix} \begin{pmatrix} h \\ k \end{pmatrix} \quad \varepsilon(\mathbf{h}) = \frac{h^2}{\|\mathbf{h}\|} = \frac{|h|}{\sqrt{1 + \frac{k^2}{h^2}}}$$

i.e. L is the linear operator with 1×2 -matrix representation $\begin{pmatrix} 4 & 1 \end{pmatrix}$ (w.r.t. the standard bases). As it is easy to see that $\varepsilon(\mathbf{h}) \rightarrow 0$ as $\mathbf{h} \rightarrow 0$, we conclude that $Df(\mathbf{x}_0) = L$. Thus f is differentiable at the point $(2, 1)$, and $Df(2, 1) = \begin{pmatrix} 4 & 1 \end{pmatrix}$.

□

Proposition 7.2.8 If $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is differentiable at $\mathbf{x} \in \mathbb{R}^n$, then it is continuous at \mathbf{x} .

□

The object of the next exercise is to supply a proof:

Exercise 7.2.9 (a) Show that if $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is differentiable at x_0 , then there are constants $C > 0$ and $\delta > 0$ such that

$$\|f(x) - f(x_0)\| \leq C\|x - x_0\| \quad \text{whenever} \quad \|x - x_0\| < \delta$$

(This is called the *Lipschitz property* of a function.)

[Hint: Write

$$f(x) = f(x_0) + Df(x_0)(x - x_0) + \varepsilon_{x_0}(x)\|x - x_0\|$$

and choose $\delta > 0$ so that $\varepsilon_{x_0}(x) < 1$ when $\|x - x_0\| < \delta$. Also use that linear operators $\mathbb{R}^n \rightarrow \mathbb{R}^m$ are bounded to find K so that $\|Df(x_0)(y)\| \leq K\|y\|$ for all $y \in \mathbb{R}^n$. Put $C := K + 1$.]

(b) Now use (a) to prove Proposition 7.2.8.

□

Examples 7.2.10 1. Consider a map $f : \mathbb{R} \rightarrow \mathbb{R}$ which is differentiable at the point $x_0 \in \mathbb{R}$. Then the usual derivative $f'(x_0)$ is a real number, whereas the derivative as we have just defined it is a linear operator $Df(x_0) : \mathbb{R} \rightarrow \mathbb{R}$. It is not hard to see that

$$Df(x_0)(h) = f'(x_0) \cdot h$$

because $\lim_{h \rightarrow 0} \frac{|f(x_0+h) - f(x_0) - f'(x_0)h|}{|h|} = 0$ by definition of $f'(x_0)$. As we pointed out in Example 7.2.1.1, every linear operator $L : \mathbb{R} \rightarrow \mathbb{R}$ is of the form $L(h) = a \cdot h$ for some $a \in \mathbb{R}$, and so linear operators $L : \mathbb{R} \rightarrow \mathbb{R}$ may be identified with real numbers $a \in \mathbb{R}$.

2. Suppose that $L : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is a linear operator. Then it is differentiable, and $DL(x) = L$ for all $x \in \mathbb{R}^n$. To see that this is so, we need only show that there is a function $\varepsilon(h) \rightarrow 0$ as $h \rightarrow 0$ so that

$$L(x+h) = L(x) + L(h) + \varepsilon(h) \|h\|$$

But since L is linear, $L(x+h) = L(x) + L(h)$ so $\varepsilon = 0$ (the constant mapping) does the trick!

Thus the derivative of a linear operator, at any point, is itself. This shouldn't be surprising if you think about it in the right way: *The best linear approximation of a linear function must surely be itself.*

□

The contents of the previous example are worth stating explicitly:

Proposition 7.2.11 *Suppose that $L : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is a linear operator. Then L is differentiable, and $DL(x) = L$ for all $x \in \mathbb{R}^n$.*

7.2.2 The Chain Rule

We need to generalize the chain rule, product rule, etc. from one to higher dimensions. We begin with the most useful one: the *chain rule*.

Example 7.2.12 In one-dimensional calculus, the chain rule is stated as follows: Suppose that $y = y(x)$, $u = u(y)$ are real-valued functions of one variable. If y is differentiable w.r.t. x and u is differentiable w.r.t. y , then u is differentiable w.r.t. x , and

$$\frac{du}{dx} = \frac{du}{dy} \frac{dy}{dx}$$

Hopefully, you have by now matured enough (mathematically) to find this confusingly imprecise. Let's try again: If $y(x)$ is differentiable at x_0 , and $u(y)$ is differentiable at $y_0 = y(x_0)$, then u is differentiable w.r.t. x at x_0 , and

$$\left. \frac{du}{dx} \right|_{x_0} = \left. \frac{du}{dy} \right|_{y_0} \left. \frac{dy}{dx} \right|_{x_0} \quad (*)$$

That's certainly better, but it's still not entirely clear. What's missing is the notion of *composition*: In order to differentiate u w.r.t. x , we want to regard u as a function of x , i.e. $u = u(y(x))$. In other words, the u in $\frac{du}{dx}$ is the composition $u(y(x)) = (u \circ y)(x)$, whereas the u in $\frac{du}{dy}$ is just $u(y)$ — they're not even the same function!

When we look at it like this, we see that what $(*)$ says is the following:

$$(u \circ y)'(x_0) = u'(y(x_0)) \cdot y'(x_0)$$

which is completely precise. Moreover, it admits generalization, as we now show.

□

Theorem 7.2.13 (Chain Rule) *Suppose that $U \subseteq \mathbb{R}^n$ is a neighbourhood of x_0 , and that $V \subseteq \mathbb{R}^m$ is a neighbourhood of y_0 . Suppose further that $f : U \rightarrow V$ is differentiable at x_0 and that $y_0 = f(x_0)$, and that $g : V \rightarrow \mathbb{R}^p$ is differentiable at y_0 . Then $g \circ f : U \rightarrow \mathbb{R}^p$ is differentiable at x_0 and*

$$D(g \circ f)(x_0) = Dg(f(x_0)) \circ Df(x_0)$$

Remark 7.2.14 Here $Df(x_0)$ is a linear operator $\mathbb{R}^n \rightarrow \mathbb{R}^m$, $Dg(f(x_0))$ is a linear operator $\mathbb{R}^m \rightarrow \mathbb{R}^p$, and $D(g \circ f)(x_0)$ is a linear operator $\mathbb{R}^n \rightarrow \mathbb{R}^p$. The preceding theorem says that the derivative

$$\mathbb{R}^n \xrightarrow{D(g \circ f)(x_0)} \mathbb{R}^p$$

is identical to the composition

$$\mathbb{R}^n \xrightarrow{Df(x_0)} \mathbb{R}^m \xrightarrow{Dg(f(x_0))} \mathbb{R}^p$$

i.e.

The derivative of the composition is the composition of the derivatives!

□

Proof of the Chain Rule: Let L denote the linear operator $Df(x_0) : \mathbb{R}^n \rightarrow \mathbb{R}^m$, and K denote the linear operator $Dg(f(x_0)) : \mathbb{R}^m \rightarrow \mathbb{R}^p$. Then $K \circ L : \mathbb{R}^n \rightarrow \mathbb{R}^p$ is certainly a linear operator (check this if it isn't obvious to you!). To verify that $K \circ L = D(g \circ f)(x_0)$, we must show that

$$(g \circ f)(x_0 + h) = (g \circ f)(x_0) + (K \circ L)(h) + \varepsilon(h) \|h\|$$

for some function $\varepsilon : \mathbb{R}^n \rightarrow \mathbb{R}^p$ satisfying $\lim_{h \rightarrow 0} \varepsilon(h) = 0$. We know, however, that there are functions $\varepsilon_1, \varepsilon_2$ so that

$$\begin{aligned} f(x_0 + h) &= f(x_0) + L(h) + \varepsilon_1(h) \|h\| \\ g(f(x_0) + k) &= g(f(x_0)) + K(k) + \varepsilon_2(k) \|k\| \end{aligned}$$

Thus

$$\begin{aligned} (g \circ f)(x_0 + h) &= g(f(x_0 + h)) \\ &= g(f(x_0) + L(h) + \varepsilon_1(h) \|h\|) \\ &= g(f(x_0)) + K(L(h) + \varepsilon_1(h) \|h\|) + \varepsilon_2(L(h) + \varepsilon_1(h) \|h\|) \|L(h) + \varepsilon_1(h) \|h\|\| \\ &= g(f(x_0)) + (K \circ L)(h) + \varepsilon(h) \|h\| \end{aligned}$$

where

$$\varepsilon(h) \|h\| = K(\varepsilon_1(h) \|h\|) + \varepsilon_2(L(h) + \varepsilon_1(h) \|h\|) \|L(h) + \varepsilon_1(h) \|h\|\|$$

To show that $\varepsilon(h) \rightarrow 0$ as $h \rightarrow 0$, it suffices to show that

$$K(\varepsilon_1(h) \|h\|) \rightarrow 0 \quad \text{as } h \rightarrow 0$$

and

$$\frac{\varepsilon_2(L(h) + \varepsilon_1(h) \|h\|) \|L(h) + \varepsilon_1(h) \|h\|\|}{\|h\|} \rightarrow 0 \quad \text{as } h \rightarrow 0$$

The first of these is easy: By Lemma 7.1.2 there is a constant α so that $\|K(x)\| \leq c\|x\|$ for all $x \in \mathbb{R}^m$. In particular,

$$\|K(\varepsilon_1(h))\| \leq \alpha\|\varepsilon_1(h)\| \rightarrow 0 \quad \text{as } h \rightarrow 0, \quad \text{because } \varepsilon_1(h) \rightarrow 0$$

To prove the second, let β be a constant so that $\|L(x)\| \leq \beta\|x\|$ for all $x \in \mathbb{R}^n$. Then

$$\frac{\|L(h) + \varepsilon_1(h)\| \|h\|}{\|h\|} \leq \frac{\|L(h)\|}{\|h\|} + \|\varepsilon_1(h)\| \leq \beta + \|\varepsilon_1(h)\|$$

Thus

$$\frac{\|\varepsilon_2(L(h) + \varepsilon_1(h)\| \|h\|)\|}{\|h\|} \leq \|\varepsilon_2(L(h) + \varepsilon_1(h)\| \|h\|)\| \left(\beta + \|\varepsilon_1(h)\| \right)$$

which is easily seen to converge to 0 as $h \rightarrow 0$.

+

We now have a beautiful definition of the derivative: The derivative of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ at a point $\mathbf{x} \in \mathbb{R}^n$ embodies the notion of linearization — it is, in essence, the linear transformation $\mathbb{R}^n \rightarrow \mathbb{R}^m$ which best approximates how f changes near \mathbf{x} .

Unfortunately, this doesn't tell us how to calculate it...

Recall, from linear algebra, that any linear transformation $L : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is essentially the same as a matrix. More precisely L has a $m \times n$ -matrix representation with respect to the standard bases, so that

$$L_{ji} = j^{\text{th}} \text{ component of } L(\mathbf{e}_i) \quad i = 1, \dots, n; \quad j = 1, \dots, m$$

where \mathbf{e}_i is the i^{th} standard basis vector. It follows that $Df(\mathbf{x})$ is essentially an $m \times n$ -matrix — the *Jacobian matrix* of f at \mathbf{x} :

$$Df(\mathbf{x})_{ji} = j^{\text{th}} \text{ component of } Df(\mathbf{x})(\mathbf{e}_i) \quad i = 1, \dots, n; \quad j = 1, \dots, m$$

The **question** is: How do we **calculate** these entries?

First, let's reduce the dimension of the problem, by looking at components.

7.2.3 Components

In this section, we look at functions $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$, in terms of their *components*. Every product $A_1 \times \cdots \times A_m$ of sets has associated with it the *projection mappings* π_1, \dots, π_m : π_j picks out the j^{th} component, i.e.

$$\pi_j : A_1 \times \cdots \times A_m \rightarrow A_j : (a_1, \dots, a_m) \mapsto a_j \quad \text{for } j = 1, \dots, m$$

Consider now a special case: If we equip \mathbb{R}^m with the standard basis, then every $\mathbf{x} \in \mathbb{R}^m$ has a representation $\mathbf{x} = (x_1, \dots, x_m)^{\text{tr}}$. The projection mappings from $\mathbb{R}^m = \underbrace{\mathbb{R} \times \cdots \times \mathbb{R}}_{m \text{ times}}$ to \mathbb{R} ,

are therefore given by

$$\pi_j : \mathbb{R}^m \rightarrow \mathbb{R} : \mathbf{x} \mapsto x^j \quad \text{for } j = 1, \dots, m$$

Thus

$$\mathbf{x} = \begin{pmatrix} \pi^1 \mathbf{x} \\ \vdots \\ \pi^m \mathbf{x} \end{pmatrix}$$

These π^j are clearly linear operators $\mathbb{R}^m \rightarrow \mathbb{R}$. (CHECK!!!)

The following lemma is obvious:

Lemma 7.2.15 *A sequence in \mathbb{R}^m converges if and only if each of its component sequences converges, i.e.*
 $\mathbf{x}_n \rightarrow \mathbf{x}$ *in \mathbb{R}^m if and only if for all $j = 1, \dots, m$ we have $\pi^j \mathbf{x}_n \rightarrow \pi^j \mathbf{x}$ in \mathbb{R} (i.e. $x_n^j \rightarrow x^j$).*

Proof: If $\mathbf{x}_n \rightarrow \mathbf{x}$, then $\pi^j \mathbf{x}_n \rightarrow \pi^j \mathbf{x}$, because the π^j are linear, and therefore continuous (cf. Corollary 7.1.3).

Conversely, suppose that each $x_n^j \rightarrow x^j$, and let $\varepsilon > 0$. By definition of “ $x_n^j \rightarrow x^j$ ”, it is possible to choose, for each $j = 1, \dots, m$, an $N_j \in \mathbb{N}$ such that

$$n \geq N_j \quad \implies \quad |x_n^j - x^j| < \frac{\varepsilon}{\sqrt{m}}$$

Let $N := \max\{N_1, \dots, N_m\}$. Then

$$n \geq N \quad \implies \quad |x_n^j - x^j| < \frac{\varepsilon}{\sqrt{m}} \quad \text{for all } j = 1, \dots, m \text{ simultaneously}$$

(because $n \geq N$ implies also $n \geq N_j$). Then

$$\|\mathbf{x}_n - \mathbf{x}\| = \sqrt{\sum_{j=1}^m |x_n^j - x^j|^2} < \sqrt{\sum_{j=1}^m \frac{\varepsilon^2}{m}} = \varepsilon$$

whenever $n \geq N$.

+

Suppose now that $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$. For $j = 1, \dots, m$, define $f^j : \mathbb{R}^n \rightarrow \mathbb{R}$ by $f^j := \pi^j \circ f$.

Example 7.2.16 If $f : \mathbb{R}^2 \rightarrow \mathbb{R}^3 : (x, y)^{tr} \mapsto (x + y, x^2 + y^2, -x - y^3)^{tr}$, then

$$f^1(x, y) = x + y, \quad f^2(x, y) = x^2 + y^2, \quad f^3(x, y) = -x - y^3$$

□

Thus every function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ can be thought of as a m -dimensional “vector of functions:

$$f = \begin{pmatrix} \pi^1 \circ f \\ \vdots \\ \pi^m \circ f \end{pmatrix} = \begin{pmatrix} f^1 \\ \vdots \\ f^m \end{pmatrix}$$

Proposition 7.2.17 *$f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is continuous at \mathbf{x}_0 if and only if each of the component functions $f^j : \mathbb{R}^n \rightarrow \mathbb{R}$ are continuous.*

Proof: (\Rightarrow): If f is continuous at \mathbf{x}_0 , so is the composition $f^j = \pi^j \circ f$: π^j is linear, linear operators between euclidean spaces are continuous (Corollary 7.1.3), and compositions of continuous functions are continuous.

(\Leftarrow): Suppose that each f^j is continuous at \mathbf{x}_0 , for $j = 1, \dots, m$. To prove f is continuous at \mathbf{x}_0 it suffices to show that if $\mathbf{x}_n \rightarrow \mathbf{x}_0$, then $f(\mathbf{x}_n) \rightarrow f(\mathbf{x}_0)$. Now we know that $f^j(\mathbf{x}_n) \rightarrow f^j(\mathbf{x}_0)$ for all $j = 1, \dots, m$.

—

7.2.4 Partial Derivatives and the Jacobian Matrix

In subsection 7.2.3 we saw that any function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ can be written as a vector of real-valued functions $f = (f^1, \dots, f^m)^{tr}$, where $f^j : \mathbb{R}^n \rightarrow \mathbb{R}$ is defined by $f^j = \pi^j \circ f$ for $j = 1, \dots, m$, and $\pi^j : \mathbb{R}^m \rightarrow \mathbb{R}$ is the j^{th} projection map.

Proposition 7.2.18 *$f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is differentiable at \mathbf{x} , if and only if each $f^j = \pi^j \circ f : \mathbb{R}^n \rightarrow \mathbb{R}$ is differentiable at \mathbf{x} (for $j = 1, \dots, m$). Then*

$$Df^j(\mathbf{x}) = D(\pi^j \circ f)(\mathbf{x}) = \pi^j \circ Df(\mathbf{x})$$

i.e. for all $\mathbf{h} \in \mathbb{R}^n$ we have

$$Df(\mathbf{x})(\mathbf{h}) = \begin{pmatrix} Df^1(\mathbf{x})(\mathbf{h}) \\ \vdots \\ Df^m(\mathbf{x})(\mathbf{h}) \end{pmatrix}$$

i.e. the matrix representation of $Df(\mathbf{x})$ has as j^{th} row the m -dimensional row vector which corresponds to the linear operator $Df^j(\mathbf{x}) : \mathbb{R}^n \rightarrow \mathbb{R}$.

Proof: First suppose that f is differentiable at \mathbf{x} . We can see that each $f^j = \pi^j \circ f$ is differentiable at \mathbf{x} , as

$$Df^j(\mathbf{x}) = D\pi^j(f(\mathbf{x})) \circ Df(\mathbf{x}) = \pi^j \circ Df(\mathbf{x})$$

where we used the chain rule and Proposition 7.2.11, applied to the (linear) projections π^j .

Conversely, suppose that each f^j is differentiable at \mathbf{x} , for $j = 1, \dots, m$, i.e. that

$$f^j(\mathbf{x} + \mathbf{h}) = f^j(\mathbf{x}) + Df^j(\mathbf{x})(\mathbf{h}) + \varepsilon^j(\mathbf{h})\|\mathbf{h}\| \quad \text{where } \varepsilon^j(\mathbf{h}) \rightarrow 0 \text{ as } \mathbf{h} \rightarrow 0$$

Define $L : \mathbb{R}^n \rightarrow \mathbb{R}^m$ by

$$L(\mathbf{h}) := \begin{pmatrix} Df^1(\mathbf{x})(\mathbf{h}) \\ \vdots \\ Df^m(\mathbf{x})(\mathbf{h}) \end{pmatrix}$$

It is clear that L is a linear operator. To see that $L = Df(\mathbf{x})$ we must show that

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + L(\mathbf{h}) + \varepsilon(\mathbf{h})\|\mathbf{h}\|$$

where $\varepsilon(\mathbf{h}) \rightarrow 0$ as $\mathbf{h} \rightarrow 0$. Now

$$f(\mathbf{x} + \mathbf{h}) = \begin{pmatrix} f^1(\mathbf{x} + \mathbf{h}) \\ \vdots \\ f^m(\mathbf{x} + \mathbf{h}) \end{pmatrix} = \begin{pmatrix} f^1(\mathbf{x}) + Df^1(\mathbf{x})(\mathbf{h}) + \varepsilon^1(\mathbf{h})\|\mathbf{h}\| \\ \vdots \\ f^m(\mathbf{x}) + Df^m(\mathbf{x})(\mathbf{h}) + \varepsilon^m(\mathbf{h})\|\mathbf{h}\| \end{pmatrix} = f(\mathbf{x}) + L(\mathbf{h}) + \varepsilon(\mathbf{h})\|\mathbf{h}\|$$

where we define $\varepsilon(\mathbf{h}) := (\varepsilon^1(\mathbf{h}), \dots, \varepsilon^m(\mathbf{h}))^{tr}$. Clearly, $\varepsilon(\mathbf{h}) \rightarrow 0$ in \mathbb{R}^m iff each $\varepsilon^j(\mathbf{h}) \rightarrow 0$ in \mathbb{R} . It follows immediately that f is differentiable at \mathbf{x} and that $Df(\mathbf{x}) = L$, as asserted.

—

Armed with proposition 7.2.18, we now have reduced the problem of finding the entries of the matrix $Df(\mathbf{x})$ to that of finding the entries of the vectors $Df^j(\mathbf{x})$:

$$Df(\mathbf{x})_{ji} = j^{\text{th}} \text{ component of } Df(\mathbf{x})(\mathbf{e}_i) = Df^j(\mathbf{x})(\mathbf{e}_i) \quad i = 1, \dots, n; \quad j = 1, \dots, m$$

To calculate the entries of the matrix $Df(\mathbf{x})$, we therefore need to investigate $Df^j(\mathbf{x})(\mathbf{e}_i)$.

Now each f^j is a map from \mathbb{R}^n to \mathbb{R} . Consider, therefore, a function $g : \mathbb{R}^n \rightarrow \mathbb{R}$ which is differentiable at \mathbf{x} . By definition,

$$g(\mathbf{x} + t\mathbf{e}_i) = g(\mathbf{x}) + Dg(\mathbf{x})(t\mathbf{e}_i) + \varepsilon(t\mathbf{e}_i)\|t\mathbf{e}_i\|$$

where $\varepsilon(\mathbf{h}) \rightarrow 0$ as $\mathbf{h} \rightarrow 0$. By linearity and some trivial calculations, this reduces to

$$g(\mathbf{x} + t\mathbf{e}_i) = g(\mathbf{x}) + t Dg(\mathbf{x})(\mathbf{e}_i) + |t| \varepsilon(t\mathbf{e}_i)$$

Rearranging, we obtain

$$\frac{g(\mathbf{x} + t\mathbf{e}_i) - g(\mathbf{x})}{t} = Dg(\mathbf{x})(\mathbf{e}_i) + \varepsilon(t\mathbf{e}_i)$$

where $\varepsilon(t\mathbf{e}_i) \rightarrow 0$ as $t \rightarrow 0$ (because then $t\mathbf{e}_i \rightarrow 0$). Taking limits, we see that

$$\lim_{t \rightarrow 0} \frac{g(\mathbf{x} + t\mathbf{e}_i) - g(\mathbf{x})}{t} = Dg(\mathbf{x})(\mathbf{e}_i)$$

i.e. that

$$Dg(\mathbf{x})(\mathbf{e}_i) = \lim_{t \rightarrow 0} \frac{g(x^1, \dots, x^i + t, \dots, x^n) - g(x^1, \dots, x^n)}{t}$$

The limit on the right ought to be very familiar: by definition

$$\lim_{t \rightarrow 0} \frac{g(x^1, \dots, x^i + t, \dots, x^n) - g(x^1, \dots, x^n)}{t} =: \left. \frac{\partial g}{\partial x^i} \right|_{\mathbf{x}}$$

i.e.

$$Dg(\mathbf{x})(\mathbf{e}_i) = \left. \frac{\partial g}{\partial x^i} \right|_{\mathbf{x}}$$

Thus:

Proposition 7.2.19 *If $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is differentiable at \mathbf{x} , then*

$$Df(\mathbf{x}) = \left(\left. \frac{\partial f}{\partial x^1} \right|_{\mathbf{x}} \quad \left. \frac{\partial f}{\partial x^2} \right|_{\mathbf{x}} \quad \dots \quad \left. \frac{\partial f}{\partial x^n} \right|_{\mathbf{x}} \right) = (\partial_1 f(\mathbf{x}) \quad \partial_2 f(\mathbf{x}) \quad \dots \quad \partial f(\mathbf{x}))$$

(i.e. $Df(\mathbf{x})$ is just the gradient $\text{grad } f = \nabla f$, but without the commas).

It follows immediately that

$$Df(\mathbf{x})_{ji} = Df^j(\mathbf{x})(\mathbf{e}_i) = \left. \frac{\partial f^j}{\partial x^i} \right|_{\mathbf{x}} = \partial_i f^j(\mathbf{x})$$

and thus that:

Theorem 7.2.20 Suppose that $f : U \rightarrow \mathbb{R}^m$, where $U \subseteq \mathbb{R}^n$ is a neighbourhood of \mathbf{x} . Then each $\frac{\partial f^j}{\partial x^i}$ exists at \mathbf{x} (for $i = 1, \dots, n$ and $j = 1, \dots, m$) and the linear operator $Df(\mathbf{x})$ has an $m \times n$ -matrix representation (w.r.t. the standard bases)

$$Df(\mathbf{x}) = \begin{pmatrix} \frac{\partial f^1}{\partial x^1} & \frac{\partial f^1}{\partial x^2} & \cdots & \frac{\partial f^1}{\partial x^n} \\ \frac{\partial f^2}{\partial x^1} & \frac{\partial f^2}{\partial x^2} & \cdots & \frac{\partial f^2}{\partial x^n} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial f^m}{\partial x^1} & \frac{\partial f^m}{\partial x^2} & \cdots & \frac{\partial f^m}{\partial x^n} \end{pmatrix}$$

where the partial derivatives are evaluated at the point \mathbf{x} .

7.2.5 A Sufficient Condition for the Existence of $Df(\mathbf{x})$

Later, we will encounter an exercise that shows that the converse of Theorem 7.2.20 is not true: There are functions f which are not differentiable at a point \mathbf{x} , but for which *all* the partial derivatives of f exist at \mathbf{x} . If we impose some mild continuity conditions on the partial derivatives, however then a converse of Theorem 7.2.20 is true. To prove it, we will need the *Mean Value Theorem* from elementary calculus. In fact, the Mean Value Theorem is one of the most important results in analysis, and we shall make heavy use of it later, when we prove Taylor's Theorem. It states that:

Theorem 7.2.21 If $f : [a, b] \rightarrow \mathbb{R}$ is differentiable on (a, b) , then there is $c \in (a, b)$ so that

$$f(b) - f(a) = f'(c)(b - a)$$

We revise the proof in the next exercise:

Exercise 7.2.22 (a) First prove *Rolle's Theorem*: If $\varphi : [a, b] \rightarrow \mathbb{R}$ is differentiable on (a, b) and $\varphi(a) = 0 = \varphi(b)$, then there is $c \in (a, b)$ so that $\varphi'(c) = 0$.

[Hint: φ must have a maximum or a minimum on $[a, b]$. At least one of these extrema must occur at an interior point $c \in (a, b)$.]

(b) Now prove the Mean Value Theorem by applying Rolle's Theorem to

$$\varphi(x) := f(x) - f(a) - \frac{f(b) - f(a)}{b - a}(x - a)$$

□

Definition 7.2.23 Let $U \subseteq \mathbb{R}^n$ be a neighbourhood of \mathbf{x} . A function $f : U \rightarrow \mathbb{R}^m$ is said to be continuously differentiable at \mathbf{x} if and only if there is an open neighbourhood $V \subseteq U$ of \mathbf{x} such that:

- (i) all the partial derivatives $\partial_i f^j$ exist in V (for $i = 1, \dots, n$ and $j = 1, \dots, m$);
- (ii) all the partial derivatives $\partial_i f^j$ are continuous at \mathbf{x} .

Theorem 7.2.24 Suppose that $U \subseteq \mathbb{R}^n$ is a neighbourhood of \mathbf{x} . If $f : U \rightarrow \mathbb{R}^m$ is continuously differentiable at \mathbf{x} , then f is differentiable at \mathbf{x} .

Proof: By Proposition 7.2.18, it suffices to prove the result for functions $f : \mathbb{R}^n \rightarrow \mathbb{R}$. Consider now the telescoping expansion

$$\begin{aligned} & f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) \\ &= f(x^1 + h^1, x^2, x^3, \dots, x^n) - f(x^1, x^2, x^3, \dots, x^n) \\ &+ f(x^1 + h^1, x^2 + h^2, x^3, \dots, x^n) - f(x^1 + h^1, x^2, x^3, \dots, x^n) \\ &+ \dots \dots \dots \\ &+ f(x^1 + h^1, x^2 + h^2, \dots, x^{n-1} + h^{n-1}, x^n + h^n) - f(x^1 + h^1, x^2 + h^2, \dots, x^{n-1} + h^{n-1}, x^n) \end{aligned}$$

Let $g_1 : \mathbb{R} \rightarrow \mathbb{R} : y \mapsto f(y, x^2, x^3, \dots, x^n)$. Then $g'_1 = \partial_1 f$. By the Mean Value Theorem, there is c_1 between x^1 and $x^1 + h^1$ so that $g'_1(b_1)h^1 = g_1(x^1 + h^1) - g_1(x^1)$, i.e. so that

$$f(x^1 + h^1, x^2, \dots, x^n) - f(x^1, x^2, \dots, x^n) = \partial_1 f(\mathbf{c}_1) \quad \text{for } \mathbf{c}_1 := (b_1, x^2, x^3, \dots, x^n)$$

Define $g_2 : \mathbb{R} \rightarrow \mathbb{R} : y \mapsto f(x^1 + h^1, y, x^3, \dots, x^n)$. Then $g'_2 = \partial_2 f$. By the Mean value Theorem again, there is b_2 between x^2 and $x^2 + h^2$ so that $g'_2(b_2)h^2 = g_2(x^2 + h^2) - g_2(x^2)$, i.e. so that

$$f(x^1 + h^1, x^2 + h^2, x^3, \dots, x^n) - f(x^1 + h^1, x^2, x^3, \dots, x^n) = \partial_2 f(\mathbf{c}_2)h^2 \quad \text{for } \mathbf{c}_2 := (x^1 + h^1, b_2, x^3, \dots, x^n)$$

Continuing in this way we see that there is $\mathbf{c}_i = (x^1 + h^1, \dots, x^{i-1} + h^{i-1}, b_i, x^{i+1}, \dots, x^n)$, where b_i lies between x^i and $x^i + h^i$, so that the i^{th} row of the telescoping expansion of $f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x})$ is $\partial_i f(\mathbf{c}_i)h^i$. Define the linear operator $L : \mathbb{R}^n \rightarrow \mathbb{R}^m$ by

$$L(\mathbf{h}) := \sum_{i=1}^n \partial_i f(\mathbf{x}) h^i$$

Then

$$f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) = \sum_{i=1}^n \partial_i f(\mathbf{c}_i) h^i = L(\mathbf{h}) + \varepsilon(\mathbf{h}) \|\mathbf{h}\|$$

where

$$\varepsilon(\mathbf{h}) = \frac{1}{\|\mathbf{h}\|} \sum_{i=1}^n \left(\partial_i f(\mathbf{c}_i) - \partial_i f(\mathbf{x}) \right) h^i$$

But

$$\|\varepsilon(\mathbf{h})\| = \frac{1}{\|\mathbf{h}\|} \left\| \sum_{i=1}^n \left(\partial_i f(\mathbf{c}_i) - \partial_i f(\mathbf{x}) \right) h^i \right\| \leq \sum_{i=1}^n \|\partial_i f(\mathbf{c}_i) - \partial_i f(\mathbf{x})\|$$

Since $\mathbf{c}_i \rightarrow \mathbf{x}$ as $\mathbf{h} \rightarrow 0$, and since each $\partial_i f$ is continuous at \mathbf{x} , we have $\partial_i f(\mathbf{c}_i) \rightarrow \partial_i f(\mathbf{x})$ as $\mathbf{h} \rightarrow 0$. It follows that $\varepsilon(\mathbf{h}) \rightarrow 0$ as $\mathbf{h} \rightarrow 0$, and thus that $L = Df(\mathbf{x})$.

◄

Exercise 7.2.25 Consider the function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ defined by

$$f(x, y) := \begin{cases} \frac{x^2 y^2}{x^4 + y^4} & \text{if } (x, y) \neq (0, 0) \\ 0 & \text{if } (x, y) = (0, 0) \end{cases}$$

- (a) Show that f is differentiable (i.e. that $Df(x, y)$ exists) at all points $(x, y) \neq (0, 0)$.
- (b) Show that f is not differentiable at $(0, 0)$.

□

7.2.6 The Chain Rule: Reprise

Let's have another look at the chain rule: Recall (or verify, if you don't recall) that if $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and $S : \mathbb{R}^m \rightarrow \mathbb{R}^p$ are linear operators, then so is their composition

$$S \circ T : \mathbb{R}^n \rightarrow \mathbb{R}^p : \mathbf{x} \mapsto S(T(\mathbf{x}))$$

The matrix representation of $S \circ T$ is the product of the matrix representations of S and T . To see this, denote for the moment the matrix representation of T (w.r.t. the standard bases) by $[T]$, so that $T(\mathbf{x})$ is the matrix \times vector $[T]\mathbf{x}$, etc. Then

$$(S \circ T)(\mathbf{x}) = [S \circ T]\mathbf{x}$$

but also

$$(S \circ T)(\mathbf{x}) = S(T(\mathbf{x})) = [S][T]\mathbf{x}$$

so that $[S \circ T] = [S][T]$, as asserted.)

We apply this result to a composition: Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and $g : \mathbb{R}^m \rightarrow \mathbb{R}^p$ be differentiable. Then

$$D(g \circ f)(\mathbf{x}) = Dg(f(\mathbf{x})) \circ Df(\mathbf{x})$$

i.e.

$$[D(g \circ f)(\mathbf{x})] = [Dg(f(\mathbf{x}))][Df(\mathbf{x})]$$

Thus:

The derivative of a composition is the product of the derivatives!

It follows that for $i = 1, \dots, n$ and $k = 1, \dots, p$ we have

$$\begin{aligned} \partial_i(g \circ f)^k(\mathbf{x}) &= [D(g \circ f)(\mathbf{x})]_{ki} \\ &= \sum_{j=1}^m [Dg(f(\mathbf{x}))]_{kj} [Df(\mathbf{x})]_{ji} \\ &= \sum_{j=1}^m \partial_j g^k(f(\mathbf{x})) \partial_i f^j(\mathbf{x}) \end{aligned}$$

In particular:

Proposition 7.2.26 *If $f = f^1, \dots, f^m : \mathbb{R}^n \rightarrow \mathbb{R}^m$ are continuously differentiable at \mathbf{x} and $g : \mathbb{R}^m \rightarrow \mathbb{R}$ is differentiable at $(f^1(\mathbf{x}), \dots, f^m(\mathbf{x}))^{tr}$, then $g \circ f$ is differentiable at \mathbf{x} and*

$$\partial_i(g \circ f)(\mathbf{x}) = \sum_{j=1}^m \partial_j g(f^1(\mathbf{x}), \dots, f^m(\mathbf{x})) \partial_i f^j(\mathbf{x})$$

In a course on advanced calculus, this is usually (and sometimes usefully) phrased as follows: Suppose that $y^j = f^j(\mathbf{x})$ are functions $\mathbb{R}^n \rightarrow \mathbb{R}^m$ which continuously differentiable at \mathbf{x} for $j = 1, \dots, m$ and that $g(y^1, \dots, y^m)$ is a function $\mathbb{R}^m \rightarrow \mathbb{R}$ which is differentiable at $\mathbf{y} = (f^1(\mathbf{x}), \dots, f^m(\mathbf{x}))^{tr}$. Then $g(y^1(\mathbf{x}), \dots, y^m(\mathbf{x}))$ is differentiable at \mathbf{x} , and

$$\frac{\partial g}{\partial x^i} = \sum_{j=1}^m \frac{\partial g}{\partial y^j} \frac{\partial y^j}{\partial x^i}$$

7.2.7 Further Manipulation Rules

Here's a nice proof that the derivative of a sum equals the sum of the derivatives:

Proposition 7.2.27 *If $f, g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ are differentiable at x , then $D(f + g)(x) = Df(x) + Dg(x)$.*

Proof: Note that $f + g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is the composition of the function $h : \mathbb{R}^n \rightarrow \mathbb{R}^m \times \mathbb{R}^m$ with the function $s : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}^m$, where

$$h(x) := \begin{pmatrix} f(x) \\ g(x) \end{pmatrix} \quad \text{and} \quad s \begin{pmatrix} x \\ y \end{pmatrix} := x + y$$

i.e. $(f+g)(x) = (s \circ h)(x)$. Now s is linear, so that $Ds(h(x)) = s$. Furthermore, by Proposition 7.2.18, we see that $Dh(x) = \begin{pmatrix} Df(x) \\ Dg(x) \end{pmatrix}$. By the chain rule

$$D(f + g)(x) = Ds(h(x)) \circ Dh(x) = s \begin{pmatrix} Df(x) \\ Dg(x) \end{pmatrix} = Df(x) + Dg(x)$$

Exercise 7.2.28 Prove Proposition 7.2.27 directly from the definition of the derivative.

□

[Henceforth, we will not always distinguish between row vectors and column vectors. Thus $Df(x, y)$ will mean the same thing as $Df \begin{pmatrix} x \\ y \end{pmatrix}$.]

Next, we tackle the multidimensional analogue of Leibniz' Rule (or the Product Rule) for differentiation. First, we need a lemma:

Lemma 7.2.29 *If $p : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}$ is defined by*

$$p(x, y) = \langle x, y \rangle$$

(where $\langle \cdot, \cdot \rangle$ is the standard inner product on \mathbb{R}^m .) Then p is differentiable, and

$$Dp(x, y)(h, k) = \langle x, k \rangle + \langle y, h \rangle$$

Proof: Fix $(x, y) \in \mathbb{R}^m \times \mathbb{R}^m$. The map $L : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}$ defined by

$$L(h, k) := \langle x, k \rangle + \langle y, h \rangle$$

is easily seen to be linear:

$$\begin{aligned} L(\alpha_1(h_1, k_1) + \alpha_2(h_2, k_2)) &= L(\alpha_1 h_1 + \alpha_2 h_2, \alpha_1 k_1 + \alpha_2 k_2) \\ &= \langle x, \alpha_1 k_1 + \alpha_2 k_2 \rangle + \langle y, \alpha_1 h_1 + \alpha_2 h_2 \rangle \\ &= \alpha_1 \langle x, k_1 \rangle + \alpha_2 \langle x, k_2 \rangle + \alpha_1 \langle y, h_1 \rangle + \alpha_2 \langle y, h_2 \rangle \\ &= \alpha_1 L(h_1, k_1) + \alpha_2 L(h_2, k_2) \end{aligned}$$

Now

$$\begin{aligned} p(x + h, y + k) &= \langle x + h, y + k \rangle \\ &= \langle x, y \rangle + \langle x, k \rangle + \langle y, h \rangle + \langle h, k \rangle \\ &= p(x, y) + L(h, k) + \varepsilon(h, k) \|(h, k)\| \end{aligned}$$

where

$$\varepsilon(h, k) := \frac{\langle h, k \rangle}{\|(h, k)\|}$$

Now $\|(h, k)\|^2 = \sum_{j=1}^m [(h^j)^2 + (k^j)^2] = \|h\|^2 + \|k\|^2 \geq \|h\|^2$. By the Cauchy–Schwarz inequality $|\langle h, k \rangle| \leq \|h\| \|k\|$, and so

$$|\varepsilon(h, k)| \leq \frac{\|h\| \|k\|}{\|h\|} = \|k\|$$

and hence $\varepsilon(h, k) \rightarrow 0$ as $(h, k) \rightarrow 0$. Hence $L = Dp(x, y)$, as required.

+

Proposition 7.2.30 *If $f, g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ are differentiable at x , then $\langle f, g \rangle : \mathbb{R}^n \rightarrow \mathbb{R}$ is differentiable at x , and*

$$D\langle f, g \rangle(x) = \langle f(x), Dg(x) \rangle + \langle g(x), Df(x) \rangle$$

i.e.

$$D\langle f, g \rangle(x)(h) = \langle Df(x)(h), g(x) \rangle + \langle f(x), Dg(x)(h) \rangle$$

Proof: If $q : \mathbb{R}^n \rightarrow \mathbb{R}^m \times \mathbb{R}^m$ and $p : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}$ are defined by

$$q(x) := (f(x), g(x)) \quad p(x, y) = \langle x, y \rangle$$

then $\langle f, g \rangle = p \circ q$. Now $Dq(x) = (Df(x), Dg(x))$ by Proposition 7.2.18, and $Dp(x, y)(h, k) = \langle x, k \rangle + \langle y, h \rangle$ by the preceding lemma. Thus, by the chain rule, we have

$$\begin{aligned} D\langle f, g \rangle(x)(h) &= D(p \circ q)(x) \\ &= Dp(q(x)) \circ Dq(x)(h) \\ &= Dp(f(x), g(x))(Df(x)(h), Dg(x)(h)) \\ &= \langle f(x), Dg(x)(h) \rangle + \langle g(x), Df(x)(h) \rangle \end{aligned}$$

+

The inner product in \mathbb{R}^1 is just multiplication: $\langle x, y \rangle = xy$. Hence:

Corollary 7.2.31 *If $f, g : \mathbb{R}^n \rightarrow \mathbb{R}$ are differentiable at x , then*

$$D(f \cdot g)(x) = f(x)Dg(x) + g(x)Df(x)$$

7.2.8 Directional Derivatives

Definition 7.2.32 Let $U \subseteq \mathbb{R}^n$ be a neighbourhood of \mathbf{x}_0 , and let $f : U \rightarrow \mathbb{R}$. If $\mathbf{v} \in \mathbb{R}^n$, then the limit (if it exists)

$$D_{\mathbf{v}}f(\mathbf{x}_0) := \lim_{t \rightarrow 0} \frac{f(\mathbf{x}_0 + t\mathbf{v}) - f(\mathbf{x}_0)}{t}$$

is called the *directional derivative* (or *Gâteaux derivative*) of f at \mathbf{x}_0 in the direction \mathbf{v} . (Here, the limit is taken over $t \in \mathbb{R}$.)

Thus: If \mathbf{v} is a unit vector, then $D_{\mathbf{v}}f(\mathbf{x}_0)$ is the instantaneous rate of change of the function f in the direction \mathbf{v} at the point \mathbf{x}_0 .

Remarks 7.2.33 1. Note that

$$D_{\mathbf{v}}f(\mathbf{x}_0) = \left. \frac{d}{dt} f(\mathbf{x}_0 + t\mathbf{v}) \right|_{t=0}$$

(If you don't see this, just define $g(t) := f(\mathbf{x}_0 + t\mathbf{v})$ and write down the definition of the derivative $g'(0)$.)

2. Also note that if \mathbf{e}_i is the i^{th} standard basis vector, then

$$D_{\mathbf{e}_i}f(\mathbf{x}_0) = \partial_i f(\mathbf{x}_0)$$

i.e. that the i^{th} partial derivative is just the directional derivative in the direction of the i^{th} basis vector.

3. If $\mathbf{0}$ is the zero vector, then $D_{\mathbf{v}}f(\mathbf{x}_0)$ is uninteresting, always being equal to zero. So when we talk about directional derivatives, we implicitly assume that the direction vector \mathbf{v} is not the zero vector.

□

The next proposition shows how directional derivatives may be obtained from the (Fréchet) derivative $Df(\mathbf{x}_0)$:

Proposition 7.2.34 Suppose that $U \subseteq \mathbb{R}^n$ is a neighbourhood of \mathbf{x}_0 , and that $f : U \rightarrow \mathbb{R}$. If f is differentiable at \mathbf{x}_0 , then all the directional derivatives $D_{\mathbf{v}}f(\mathbf{x}_0)$ exist (for all $\mathbf{v} \in \mathbb{R}^n$ with $\mathbf{v} \neq \mathbf{0}$), and

$$D_{\mathbf{v}}f(\mathbf{x}_0) = Df(\mathbf{x}_0)(\mathbf{v})$$

Proof: By definition of the Fréchet derivative, we have

$$f(\mathbf{x}_0 + t\mathbf{v}) = f(\mathbf{x}_0) + tDf(\mathbf{x}_0)(\mathbf{v}) + \varepsilon(t\mathbf{v}) \|\mathbf{v}\|$$

where $\varepsilon(\mathbf{h}) \rightarrow 0$ as $\mathbf{h} \rightarrow 0$. Rearranging, we see that

$$\left| \frac{f(\mathbf{x}_0 + t\mathbf{v}) - f(\mathbf{x}_0)}{t} - Df(\mathbf{x}_0)(\mathbf{v}) \right| = |\varepsilon(t\mathbf{v})| \|\mathbf{v}\| \rightarrow 0 \quad \text{as } t \rightarrow 0$$

Hence $\lim_{t \rightarrow 0} \frac{f(\mathbf{x}_0 + t\mathbf{v}) - f(\mathbf{x}_0)}{t} = Df(\mathbf{x}_0)(\mathbf{v})$, as asserted.

□

Remarks 7.2.35 Consider a function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$. The graph of f is a surface $z = f(\mathbf{x})$ in \mathbb{R}^3 (where $\mathbf{x} = (x, y)$), and the point $\mathbf{x}_0, f(\mathbf{x}_0)$ lies on the graph. If \mathbf{v} is a non-zero vector in \mathbb{R}^2 , the instantaneous rate of change in the direction of \mathbf{v} is $D_{\mathbf{v}}f(\mathbf{x}_0)$, i.e.

$$f(\mathbf{x}_0 + t\mathbf{v}) \approx f(\mathbf{x}_0) + tD_{\mathbf{v}}f(\mathbf{x}_0)$$

for small t . So

$$z = f(\mathbf{x}_0) + tD_{\mathbf{v}}f(\mathbf{x}_0) \quad t \in \mathbb{R}$$

gives the tangent to the surface at the point \mathbf{x}_0 in the direction \mathbf{v} . The collection of all these tangent lines forms the tangent plane, and the tangent plane may therefore be described by the equation

$$z = f(\mathbf{x}_0) + Df(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)$$

i.e.

$$z = f(x_0, y_0) + \partial_1 f(x_0, y_0) \cdot (x - x_0) + \partial_2 f(x_0, y_0) \cdot (y - y_0)$$

□

Note, however that the converse of Proposition 7.2.34 is not true: It is possible for all directional derivatives of f to exist at a point \mathbf{x}_0 , and yet for f not to be differentiable at \mathbf{x}_0 . In particular, it is possible for all partial derivatives of f to exist at \mathbf{x}_0 , and yet for f to be non-differentiable at \mathbf{x}_0 . This is the content of the next example:

Example 7.2.36 Consider the function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ defined by

$$f(x, y) := \begin{cases} \frac{xy}{x^2 + y} & \text{if } x^2 \neq -y \\ 0 & \text{if } x^2 = -y \end{cases}$$

If $\mathbf{v} = (v_1, v_2)$, then

$$\frac{f(\mathbf{0} + t\mathbf{v}) - f(\mathbf{0})}{t} = \frac{1}{t} \frac{t^2 v_1 v_2}{t^2 v_1 + t v_2} = \frac{v_1 v_2}{t v_1 + v_2}$$

Taking the limit as $t \rightarrow 0$ we see that

$$D_{\mathbf{v}} f(\mathbf{0}) = \begin{cases} v_1 & \text{if } v_2 \neq 0 \\ 0 & \text{if } v_2 = 0 \end{cases}$$

Hence all directional derivatives exist at $\mathbf{0}$.

However, f is not continuous at $\mathbf{0}$, and therefore not differentiable there: For example $\mathbf{x}_n = (\frac{1}{n}, \frac{1}{\sqrt{n}})$ has $\mathbf{x}_n \rightarrow \mathbf{0}$. Yet

$$f(\mathbf{x}_n) = \frac{\frac{1}{n\sqrt{n}}}{\frac{1}{n^2} + \frac{1}{\sqrt{n}}} = \frac{\sqrt{n}}{1 + n^{\frac{3}{2}}} \rightarrow \infty \quad \text{as } n \rightarrow \infty$$

□

Exercise 7.2.37 Let $A \subseteq \mathbb{R}^2$ be given by

$$A := \{(x, y) : x > 0 \text{ and } 0 < y < x^2\}$$

Define $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ by

$$f(x, y) = \begin{cases} 1 & \text{if } (x, y) \in A \\ 0 & \text{else} \end{cases}$$

- (a) Show that $D_{\mathbf{v}} f(0, 0) = 0$ for all $\mathbf{v} \in \mathbb{R}^2$.
- (b) Show that f is not differentiable at $(0, 0)$.

[Hints: (a) First show that any straight line through $(0, 0)$ contains a line segment about $(0, 0)$ which lies wholly in the complement of A . Then show that for any $\mathbf{v} \neq \mathbf{0}$ there is $\delta > 0$ so that $f(t\mathbf{v}) = 0$ whenever $|t| < \delta$.

Show that f is not continuous at $(0, 0)$.]

□

7.3 Taylor Theorems

7.3.1 Taylor's Theorem for One Variable

We introduce a number of Taylor Theorems in one dimension. As these are probably already familiar, we omit discussion of these results.

Before we continue, recall the Mean Value Theorem, which is Theorem 7.2.21.

Theorem 7.3.1 (Taylor's Theorem)

Let $n \in \mathbb{N}$. Suppose that $f : [a, b] \rightarrow \mathbb{R}$ is a function with the property that

- (i) $f, f', f'', \dots, f^{(n-1)}$ are defined and continuous on $[a, b]$;
- (ii) $f^{(n)}$ exists on (a, b) .

Suppose further that α, β are real numbers such that $a \leq \alpha < \beta \leq b$. Then there exists a $\gamma \in (\alpha, \beta)$ such that

$$f(\beta) = \sum_{k=0}^{n-1} \frac{f^{(k)}(\alpha)}{k!} (\beta - \alpha)^k + \frac{f^{(n)}(\gamma)}{n!} (\beta - \alpha)^n$$

Proof: For $x \in [a, b]$, define

$$P(x) = \sum_{k=0}^{n-1} \frac{f^{(k)}(\alpha)}{k!} (x - \alpha)^k$$

We must show that there is a $\gamma \in (\alpha, \beta)$ such that $f(\beta) = P(\beta) + \frac{f^{(n)}(\gamma)}{n!} (\beta - \alpha)^n$. Now let $M = \frac{f(\beta) - P(\beta)}{(\beta - \alpha)^n}$, so that

$$f(\beta) = P(\beta) + M(\beta - \alpha)^n$$

and define

$$g(x) = f(x) - P(x) - M(x - \alpha)^n \quad (x \in [a, b])$$

Note that $g(\beta) = 0$, by definition of M . Note also that

$$g^{(k)}(\alpha) = f^{(k)}(\alpha) - P^{(k)}(\alpha) = 0 \quad k = 0, 1, \dots, n-1$$

because $P^{(k)}(\alpha) = f^{(k)}(\alpha)$ for such k .

We must show that $M = \frac{f^{(n)}(\gamma)}{n!}$ for some γ between α and β . Note that $P(x)$ is an $(n-1)^{\text{th}}$ degree polynomial in x , so that $P^{(n)}(x) = 0$. It follows that

$$g^{(n)}(x) = f^{(n)}(x) - n!M$$

Thus if we can find a γ such that $g^{(n)}(\gamma) = 0$, we will have $M = \frac{f^{(n)}(\gamma)}{n!}$, as required.

Now both $g(\alpha) = 0$ and $g(\beta) = 0$. By the Mean Value Theorem, there is $\gamma_1 \in (\alpha, \beta)$ such that $g'(\gamma_1) = 0$. Thus both $g'(\alpha) = 0$ and $g'(\gamma_1) = 0$. By the Mean Value Theorem, there is $\gamma_2 \in (\alpha, \gamma_1)$ such that $g''(\gamma_2) = 0$. Thus both $g''(\alpha) = 0$ and $g''(\gamma_2) = 0$. By the Mean Value Theorem, there is $\gamma_3 \dots$

After $n-1$ steps, we obtain, from the fact that both $g^{(n-1)}(\alpha) = 0$ and $g^{(n-1)}(\gamma_{n-1}) = 0$, a $\gamma_n \in (\alpha, \gamma_{n-1})$ such that $g^{(n)}(\gamma_n) = 0$.

γ_n is therefore the γ that we seek.

◄

We leave the proofs of the following results as a *very important exercise*:

Theorem 7.3.2 (Taylor's Theorem) *If $f : (x - a, x + a) \rightarrow \mathbb{R}$ is n times differentiable, with $|f^{(n)}(t)| \leq M$ for all $t \in (x - a, x + a)$, then*

$$\left| f(t) - \sum_{j=0}^{n-1} \frac{f^{(j)}(x)}{j!} (t-x)^j \right| \leq \frac{M|t-x|^n}{n!}$$

□

Theorem 7.3.3 (Taylor's Theorem) *If $f : (x - a, x + a) \rightarrow \mathbb{R}$ is n times differentiable, and $f^{(n+1)}(x)$ exists, then*

$$f(t) = \sum_{j=0}^{n+1} \frac{f^{(j)}(x)}{j!} (t-x)^j + \varepsilon(t)|t-x|^{n+1} \quad \text{where } \varepsilon(t) \rightarrow 0 \text{ as } t \rightarrow x$$

□

Exercise 7.3.4 We prove the preceding theorems. (This exercise is from *A Companion to Analysis*, by T.W. Körner, publ. Amer. Math. Soc. (2004))

Without loss of generality, we may assume that $x = 0$. Consider two functions $f, g : (-a, a) \rightarrow \mathbb{R}$, where $a > 0$.

- (a) Suppose that f, g are differentiable, that $f(0) = g(0)$, and that $f'(t) \leq g'(t)$ for all $0 \leq t < a$. Show that $f(t) \leq g(t)$ for all $0 \leq t < a$.
- (b) Suppose that $f(0) = 0$, and that $|f'(t)| \leq |t|^r$ for all $t \in (-a, a)$. Show that $|f(t)| \leq |t|^{r+1}/(r+1)$ for all $|t| < a$.
- (c) Suppose that g is n times differentiable, with $|g^{(n)}(t)| \leq M$ for all $t \in (-a, a)$, and that $g(0) = g'(0) = g''(0) = \dots = g^{(n-1)}(0) = 0$. Use (b) n times to show that

$$|g(t)| \leq \frac{M|t|^n}{n!} \quad \text{for all } |t| < a$$

- (d) Suppose that f is n times differentiable, with $|f^{(n)}(t)| \leq M$ for all $t \in (-a, a)$. Show that

$$\left| f(t) - \sum_{j=0}^{n-1} \frac{f^{(j)}(0)}{j!} t^j \right| \leq \frac{M|t|^n}{n!} \quad \text{for all } |t| < a$$

This proves Theorem 7.3.2.

[Hint: Set $g(t) := f(t) - \sum_{j=0}^{n-1} \frac{f^{(j)}(0)}{j!} t^j$ and apply (c).]

- (e) Suppose that g is n times differentiable in $(-a, a)$, and $n+1$ times differentiable at 0. Further assume that $g(0) = g'(0) = g''(0) = \dots = g^{(n+1)}(0) = 0$. Show, using (c), that

$$|g(t)| \leq \eta(t)|t| \frac{|t|^n}{n!} \quad \text{where } \eta(t) \rightarrow 0 \text{ as } t \rightarrow 0$$

[Hint: First use the definition of the derivative to explain why $g^{(n)}(u) = \varepsilon(u)u$ for some function ε such that $\varepsilon(u) \rightarrow 0$ as $u \rightarrow 0$. Then set $\eta(t) := \max_{-t \leq v \leq t} |\varepsilon(v)|$, and set $M_t := \eta(t)|t|$. Show that $|g^{(n)}(u)| \leq M_t$ for all $u \in (-t, t)$.]

- (f) If f is n times differentiable in $(-a, a)$, and $n+1$ times differentiable at 0, show that

$$\left| f(t) - \sum_{j=0}^{n+1} \frac{f^{(j)}(0)}{j!} t^j \right| \leq \frac{\eta(t)|t|^{n+1}}{n!} \quad \text{where } \eta(t) \rightarrow 0 \text{ as } t \rightarrow 0$$

This proves Theorem 7.3.3

□

7.3.2 A Higher-Order Taylor Theorem

For the one-dimensional Taylor Theorems, we employed the higher derivatives $f''(x), f^{(3)}(x), \dots, f^{(k)}(x)$ of the function $f : \mathbb{R} \rightarrow \mathbb{R}$. If $f : \mathbb{R}^n \rightarrow \mathbb{R}$, it is not yet clear what *kind of objects* the higher derivatives $D^2f(\mathbf{x}), D^3f(\mathbf{x}), \dots, D^kf(\mathbf{x})$ actually *are*. We know that $Df(\mathbf{x})$ is a linear map. It will turn out later that $D^2f(\mathbf{x})$ is a *bilinear* map, and that, in general, the $D^kf(\mathbf{x})$ are multilinear mappings from $\underbrace{\mathbb{R}^n \times \dots \times \mathbb{R}^n}_{k \text{ times}}$ to \mathbb{R} . We will attempt to do without the

use of these objects.

Remarks 7.3.5 In what follows, we will use the Mean Value Theorem over and over, in the following manner: Suppose that $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ is differentiable at the point $\mathbf{x} = (x, y)$. We will be interested in how much the function f can change in a neighbourhood of \mathbf{x} , i.e., with $\mathbf{h} = (h, k)$, we will be interested in finding bounds for the expression

$$S := |f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x})| = |f(x + h, y + k) - f(x, y)|$$

Now clearly,

$$\begin{aligned} S &= |f(x + h, y + k) - f(x, y + k) + f(x, y + k) - f(x, y)| \\ &\leq |f(x + h, y + k) - f(x, y + k)| + |f(x, y + k) - f(x, y)| \end{aligned}$$

and so we will find ourselves looking at expressions of the form $f(x, y + k) - f(x, y)$ or $f(x + h, y) - f(x, y)$. If we fix x , then $g(y) := f(x, y)$ is a function of y , and

$$g'(y) = f_{,2}(y)$$

Now $f(x, y + k) - f(x, y) = g(y + k) - g(y)$, so by the Mean value Theorem there is k^* between 0 and k so that

$$g(y + k) - g(y) = g'(y + k^*)k$$

i.e.

$$f(x, y + k) - f(x, y) = f_{,2}(x, y + k^*)k$$

Similarly, if we keep $y + k$ fixed, and apply the Mean Value Theorem to $\tilde{g}(x) := f(x, y + k)$, we see that

$$f(x + h, y + k) - f(x, y + k) = f_{,1}(x + h^*, y + k)h$$

Putting this together, we see that

$$S \leq |f_{,1}(x + h^*, y + k)| |h| + |f_{,2}(x, y + k^*)| |k|$$

Thus knowledge of the size of partial derivatives in a neighbourhood of (x, y) will give us information about the size of S , the change in f . □

We'll begin by looking at the second-order Taylor expansion of a function of two variables $f : \mathbb{R}^2 \rightarrow \mathbb{R}$. One important fact that we will state upfront is the following, though you no doubt already know it:

$$\frac{\partial^2 f}{\partial x^i \partial x^j} = \frac{\partial^2 f}{\partial x^j \partial x^i}$$

when these derivatives are continuous. To simplify notation, we define

$$f_{,i} := \frac{\partial f}{\partial x^i} = \partial_i f \quad f_{,ij} := \frac{\partial f}{\partial x^i \partial x^j} = \partial_i \partial_j f$$

etc.

Theorem 7.3.6 Suppose that $f : U \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$ is twice differentiable on the open set U , and that the functions $\frac{\partial^2 f}{\partial x^i \partial x^j}$ are continuous. Then

$$\frac{\partial^2 f}{\partial x^i \partial x^j}$$

Proof: ¹ For simplicity, by holding all other variables fixed, we may assume that $f : U \subseteq \mathbb{R}^2 \rightarrow \mathbb{R}$. Let $\mathbf{x} = (x, y) \in U \subseteq \mathbb{R}^2$, and let $\mathbf{h} = (h, k)$. Consider

$$S_{h,k} := [f(x+h, y+k) - f(x+h, y)] - [f(x, y+k) - f(x, y)]$$

If we define

$$g_k(x) := f(x, y+k) - f(x, y)$$

we see that

$$S_{h,k} = g_k(x+h) - g_k(x)$$

By the Mean Value Theorem there is $c_{k,h} \in (x, x+h)$ so that

$$S_{h,k} = g'_k(c_{k,h}) \cdot h$$

and hence

$$S_{h,k} = \left[\frac{\partial f}{\partial x} \Big|_{(c_{k,h}, y+k)} - \frac{\partial f}{\partial x} \Big|_{(c_{k,h}, y)} \right] \cdot h$$

By the Mean Value Theorem, again, we see that

$$S_{h,k} = \frac{\partial^2 f}{\partial y \partial x} \Big|_{(c_{k,h}, d_{k,h})} \cdot hk$$

for some $d_{k,h} \in (y, y+k)$.

Now note that also

$$S_{h,k} = [f(x+h, y+k) - f(x, y+k)] - [f(x+h, y) - f(x, y)] = \tilde{g}_h(y+k) - \tilde{g}_h(y)$$

where $\tilde{g}_h(y) := f(x+h, y) - f(x, y)$. In exactly the same way as above, we can find $\tilde{d}_{h,k} \in (y, y+k)$ and the $\tilde{c}_{h,k} \in (x, x+h)$ so that

$$S_{h,k} = \frac{\partial^2 f}{\partial x \partial y} \Big|_{(\tilde{c}_{h,k}, \tilde{d}_{h,k})} \cdot kh$$

As this is true for all h, k it remains true if we let $h, k \rightarrow 0$, by continuity of the partial derivatives. Then $c_{k,h}, \tilde{c}_{h,k} \rightarrow x$ and $d_{k,h}, \tilde{d}_{h,k} \rightarrow y$ and so

$$\frac{\partial^2 f}{\partial x \partial y} \Big|_{(x,y)} = \frac{\partial^2 f}{\partial y \partial x} \Big|_{(x,y)}$$

as claimed. +

We can now write down the promised second-order Taylor Theorem.

Theorem 7.3.7 Suppose that $U \subseteq \mathbb{R}^2$ is a neighbourhood of \mathbf{x}_0 . If the partial derivatives $\frac{\partial^2 f}{\partial x^2}, \frac{\partial^2 f}{\partial x \partial y}, \frac{\partial^2 f}{\partial y^2}$ exist in U and are continuous at $\mathbf{x} = (x, y)$, and if $\mathbf{h} = (h, k)$, then

$$f(\mathbf{x}_0 + \mathbf{h}) = f(\mathbf{x}_0) + f_{,1}(\mathbf{x}_0)h + f_{,2}(\mathbf{x}_0)k + \frac{1}{2}(f_{,11}(\mathbf{x}_0)h^2 + 2f_{,12}(\mathbf{x}_0)hk + f_{,22}(\mathbf{x}_0)k^2) + \varepsilon(\mathbf{h})\|\mathbf{h}\|^2$$

where

$$\varepsilon(\mathbf{h}) \rightarrow 0 \quad \text{as} \quad \mathbf{h} \rightarrow 0$$

¹Taken from *Elementary Classical Analysis*, by J.E. Marsden.

To prove this second-order Taylor Theorem, we need a peculiar lemma:

Lemma 7.3.8 *Suppose that $U \subseteq \mathbb{R}^2$ is a neighbourhood of 0, and that $f : U \rightarrow \mathbb{R}$ is a function whose second partial derivatives exist in U , and are continuous at 0. Suppose in addition that*

$$0 = f(0, 0) = f_{,1}(0, 0) = f_{,2}(0, 0) = f_{,11}(0, 0) = f_{,12}(0, 0) = f_{,22}(0, 0)$$

Then

$$\frac{f(\mathbf{h})}{\|\mathbf{h}\|^2} \rightarrow 0 \quad \text{as} \quad \mathbf{h} \rightarrow 0$$

Remarks 7.3.9 Suppose that f is as in the Lemma. The second order Taylor expansion of f about $(0, 0)$ is then given by

$$f(h, k) = f(0, 0) + (f_{,1}(0, 0)h + f_{,2}(0, 0)k) + \frac{1}{2}(f_{,11}(0, 0)h^2 + 2f_{,12}(0, 0)hk + f_{,22}(0, 0)k^2) + \varepsilon(\mathbf{h})\|\mathbf{h}\|^2$$

i.e.

$$f(\mathbf{h}) = \varepsilon(\mathbf{h})\|\mathbf{h}\|^2$$

as the function and all the derivatives are zero at $(0, 0)$. Thus $\varepsilon(\mathbf{h}) = \frac{f(\mathbf{h})}{\|\mathbf{h}\|^2}$. To prove that $\frac{f(\mathbf{h})}{\|\mathbf{h}\|^2} \rightarrow 0$ is thus to prove that $\varepsilon(\mathbf{h}) \rightarrow 0$ (as $\mathbf{h} \rightarrow 0$).

In other words, to prove the Lemma is to prove the 2nd order Taylor Theorem about $(0, 0)$ for functions f with the properties given in the statement of the Lemma. It turns out that the 2nd order Taylor Theorem for an arbitrary function about an arbitrary point can easily be reduced to this case. So this Lemma does all the hard work. □

Proof of Lemma 7.3.8: Write $\mathbf{h} = (h, k)$. Fix $\varepsilon > 0$. Since the 2nd order partial derivatives are continuous at $(0, 0)$, there is $\delta > 0$ so that

$$|f_{,11}(\mathbf{h})|, |f_{,12}(\mathbf{h})|, |f_{,22}(\mathbf{h})| < \varepsilon \quad \text{whenever } \|\mathbf{h}\| < \delta$$

Applying the Mean Value Theorem to the function $g(k) = f_{,1}(h, k)$, we see that $g(k) - g(0) = g'(k^*)k$ for some $k^* \in (0, k)$, and thus that $f_{,1}(h, k) - f_{,1}(h, 0) = f_{,12}(h, k^*)k$ and thus that

$$|f_{,1}(h, k) - f_{,1}(h, 0)| \leq \varepsilon|k| \leq \varepsilon\|\mathbf{h}\| \quad \text{when } \|\mathbf{h}\| < \delta$$

A similar argument shows that

$$|f_{,1}(h, 0) - f_{,1}(0, 0)| \leq \varepsilon|h| \leq \varepsilon\|\mathbf{h}\| \quad \text{when } \|\mathbf{h}\| < \delta$$

Now

$$\begin{aligned} |f_{,1}(h, k)| &= |f_{,1}(h, k) - f_{,1}(0, 0)| \\ &\leq |f_{,1}(h, k) - f_{,1}(h, 0)| + |f_{,1}(h, 0) - f_{,1}(0, 0)| \\ &\leq 2\varepsilon\|\mathbf{h}\| \quad \text{when } \|\mathbf{h}\| < \delta \end{aligned} \tag{*}$$

Using the Mean Value Theorem again, we see that $f_{,2}(0, k) - f_{,2}(0, 0) = f_{,22}(0, \hat{k})k$ for some $\hat{k} \in (0, k)$, and thus that

$$|f_{,2}(0, k) - f_{,2}(0, 0)| < \varepsilon|k| \leq \varepsilon\|\mathbf{h}\| \quad \text{when } \|\mathbf{h}\| < \delta \tag{**}$$

Similarly, $f(h, k) - f(0, k) = f_{,1}(\tilde{h}, k)h$ for some $\tilde{h} \in (0, h)$. But $||(\tilde{h}, k)|| \leq ||(h, k)||$, and thus $|f_{,1}(\tilde{h}, k)| < 2\varepsilon||\mathbf{h}||$ when $||\mathbf{h}|| < \delta$, by (*). It follows that

$$|f(h, k) - f(0, k)| < 2\varepsilon||\mathbf{h}|| \cdot |h| \leq 2\varepsilon||\mathbf{h}||^2 \quad \text{when } ||\mathbf{h}|| < \delta \quad (\dagger)$$

Similarly, from (**), we see that

$$|f(0, k) - f(0, 0)| < \varepsilon||\mathbf{h}||^2 \quad \text{when } ||\mathbf{h}|| < \delta \quad (\ddagger)$$

Putting (\dagger), (\ddagger) together, we see that

$$\begin{aligned} |f(h, k)| &= |f(h, k) - f(0, k) + f(0, k) - f(0, 0)| \\ &\leq |f(h, k) - f(0, k)| + |f(0, k) - f(0, 0)| \\ &\leq 3\varepsilon||\mathbf{h}||^2 \quad \text{when } ||\mathbf{h}|| < \delta \end{aligned}$$

and thus that

$$\frac{|f(\mathbf{h})|}{||\mathbf{h}||^2} < 3\varepsilon \quad \text{when } ||\mathbf{h}|| < \delta$$

Since ε was arbitrary, it follows that $\frac{f(\mathbf{h})}{||\mathbf{h}||^2} \rightarrow 0$ as $\mathbf{h} \rightarrow 0$.

—

Proof of Theorem 7.3.7: We may assume that $\mathbf{x}_0 = 0$, by translation². Define $g : U \rightarrow \mathbb{R}$ by

$$g(x, y) := f(x, y) - [f(0, 0) + f_{,1}(0, 0)x + f_{,2}(0, 0)y + \frac{1}{2}(f_{,11}(0, 0)x^2 + 2f_{,12}(0, 0)xy + f_{,22}(0, 0)y^2)]$$

Then

$$0 = g(0, 0) = g_{,1}(0, 0) = g_{,2}(0, 0) = g_{,11}(0, 0) = g_{,12}(0, 0) = g_{,22}(0, 0)$$

By the preceding lemma, we see that

$$\frac{g(\mathbf{h})}{||\mathbf{h}||} \rightarrow 0 \quad \text{as } ||\mathbf{h}||^2 \rightarrow 0$$

It is now clear that the Theorem holds, with $\varepsilon(\mathbf{h}) := \frac{g(\mathbf{h})}{||\mathbf{h}||^2}$ and $\mathbf{x}_0 = 0$.

—

Now that we have seen the form of a two-dimensional Taylor formula, it is not hard to generalize it to higher dimensions:

Theorem 7.3.10 Suppose that $U \subseteq \mathbb{R}^m$ is a neighbourhood of \mathbf{x}_0 , and that $f : U \rightarrow \mathbb{R}$. Suppose further that all the partial derivatives $f_{,i}, f_{,ij}, f_{,ijk}, \dots$ up to n^{th} order exist in U , and are continuous at \mathbf{x}_0 . Then

$$\begin{aligned} f(\mathbf{x}_0 + \mathbf{h}) &= f(\mathbf{x}_0) + \sum_{i=1}^m f_{,i}(\mathbf{x}_0)h^i + \frac{1}{2!} \sum_{i=1}^m \sum_{j=1}^m f_{,ij}(\mathbf{x}_0)h^i h^j \\ &\quad + \frac{1}{3!} \sum_{i=1}^m \sum_{j=1}^m \sum_{k=1}^m f_{,ijk}(\mathbf{x}_0)h^i h^j h^k + \dots \\ &\quad + \frac{1}{n!} \text{sum up to } n^{\text{th}} \text{ order} + \varepsilon(\mathbf{h})||\mathbf{h}||^n \end{aligned}$$

where $\varepsilon(\mathbf{h}) \rightarrow 0$ as $\mathbf{h} \rightarrow 0$.

²That is, define $g(\mathbf{x}) = f(\mathbf{x} + \mathbf{x}_0)$, so that $g(0) = f(\mathbf{x}_0)$. Then redefine f to be g .

Proof: We just give an outline of the proof. All the important elements are contained in the proof of Theorem 7.3.7, and the lemma that precedes it.

Consider the function

$$\begin{aligned} g(\mathbf{h}) := & f(\mathbf{x}_0 + \mathbf{h}) - \left[f(\mathbf{x}_0) + \sum_{i=1}^m f_{,i}(\mathbf{x}_0) h^i \right. \\ & + \frac{1}{2!} \sum_{i=1}^m \sum_{j=1}^m f_{,ij}(\mathbf{x}_0) h^i h^j \\ & + \frac{1}{3!} \sum_{i=1}^m \sum_{j=1}^m \sum_{k=1}^m f_{,ijk}(\mathbf{x}_0) h^i h^j h^k + \dots \\ & \left. + \frac{1}{n!} \text{sum up to } n^{\text{th}} \text{ order} \right] \end{aligned}$$

We see that $g(0)$ and all the partial derivatives $g_{,i}(0), g_{,ij}(0), g_{,ijk}(0), \dots$ up to n^{th} order are *all* equal to 0.

Now we imitate the proof of Lemma 7.3.8: Fix $\bar{\varepsilon} > 0$, and, by continuity of the n^{th} order partial derivatives at 0, choose $\bar{\delta} > 0$ so that *all* the n^{th} order partial derivatives $g_{,i_1 \dots i_n}$ have

$$|g_{,i_1 \dots i_n}(\mathbf{h})| < \bar{\varepsilon} \quad \text{when } \|\mathbf{h}\| < \bar{\delta}$$

Apply the Mean Value Theorem to show that all the $(n-1)^{\text{th}}$ derivatives satisfy

$$|g_{,j_1 \dots j_{n-1}}(\mathbf{h})| < C\bar{\varepsilon}\|\mathbf{h}\| \quad \text{when } \|\mathbf{h}\| < \bar{\delta}$$

for some constant C . Then apply the Mean Value Theorem again to show that all the $(n-2)^{\text{th}}$ derivatives satisfy

$$|g_{,k_1 \dots k_{n-2}}(\mathbf{h})| < C\bar{\varepsilon}\|\mathbf{h}\|^2 \quad \text{when } \|\mathbf{h}\| < \bar{\delta}$$

for some (different) constant C . Keep going, reducing the order of the partial derivatives step by step, until one deduces that

$$|g(\mathbf{h})| \leq C\bar{\varepsilon}\|\mathbf{h}\|^n \quad \text{when } \|\mathbf{h}\| < \bar{\delta}$$

for some constant C . We then see that

$$\varepsilon(\mathbf{h}) := \frac{|g(\mathbf{h})|}{\|\mathbf{h}\|^n} \leq C\bar{\varepsilon} \quad \text{when } \|\mathbf{h}\| < \bar{\delta}$$

Since $\bar{\varepsilon}$ was arbitrary, it follows that $\varepsilon(\mathbf{h}) \rightarrow 0$ as $\mathbf{h} \rightarrow 0$. By rearranging the definition of $g(\mathbf{h})$, we obtain the desired n^{th} order Taylor expansion of f at the point \mathbf{x}_0 .

—

Finally, applying Theorem 7.3.10 to the components of a function $f : \mathbb{R}^m \rightarrow \mathbb{R}^p$, we obtain the following straightforward extension:

Theorem 7.3.11 Suppose that $U \subseteq \mathbb{R}^m$ is a neighbourhood of \mathbf{x}_0 , and that $f = (f^1, \dots, f^l) : U \rightarrow \mathbb{R}^p$. Suppose further that all the partial derivatives up to order n exist in U and are continuous at \mathbf{x}_0 . Then for $l = 1, \dots, m$ we have

$$\begin{aligned} f^l(\mathbf{x}_0 + \mathbf{h}) = & f^l(\mathbf{x}_0) + \sum_{i=1}^m f^l_{,i}(\mathbf{x}_0)h^i + \frac{1}{2!} \sum_{i=1}^m \sum_{j=1}^m f^l_{,ij}(\mathbf{x}_0)h^i h^j \\ & + \frac{1}{3!} \sum_{i=1}^m \sum_{j=1}^m \sum_{k=1}^m f^l_{,ijk}(\mathbf{x}_0)h^i h^j h^k + \dots \\ & + \frac{1}{n!} \text{sum up to } n^{\text{th}} \text{ order} + \varepsilon^l(\mathbf{h}) \|\mathbf{h}\|^n \end{aligned}$$

where $\varepsilon^l(\mathbf{h}) \rightarrow 0$ as $\mathbf{h} \rightarrow 0$.

□

We finish this subsection by briefly revisiting the first- and second order Taylor expansions of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$. We have seen that the derivative, when it exists, is given by the Jacobian matrix:

$$J_{\mathbf{x}_0}(f) = Df(\mathbf{x}_0) = \begin{pmatrix} f_{,1}(\mathbf{x}_0) & \dots & f_{,n}(\mathbf{x}_0) \end{pmatrix}$$

The first order Taylor expansion of f about \mathbf{x}_0 is therefore given by

$$f(\mathbf{x}_0 + \mathbf{h}) = f(\mathbf{x}_0) + J_{\mathbf{x}_0}(f)\mathbf{h} + \varepsilon(\mathbf{h})\|\mathbf{h}\|$$

Define the *Hessian matrix* of f at \mathbf{x}_0 by

$$H_{\mathbf{x}_0}(f) = \begin{pmatrix} f_{,11}(\mathbf{x}_0) & f_{,12}(\mathbf{x}_0) & \dots & f_{,1n}(\mathbf{x}_0) \\ f_{,21}(\mathbf{x}_0) & f_{,22}(\mathbf{x}_0) & \dots & f_{,2n}(\mathbf{x}_0) \\ \vdots & \vdots & \dots & \vdots \\ f_{,n1}(\mathbf{x}_0) & f_{,n2}(\mathbf{x}_0) & \dots & f_{,nn}(\mathbf{x}_0) \end{pmatrix} = \begin{pmatrix} \frac{\partial^2 f}{\partial x^1 \partial x^1} & \dots & \frac{\partial^2 f}{\partial x^1 \partial x^n} \\ \vdots & \dots & \vdots \\ \frac{\partial^2 f}{\partial x^n \partial x^1} & \dots & \frac{\partial^2 f}{\partial x^n \partial x^n} \end{pmatrix} \Big|_{\mathbf{x}_0}$$

By equality of the mixed partial derivatives, the Hessian is a *symmetric matrix*. Now note that the second-order Taylor expansion of f at \mathbf{x}_0 is given by

$$f(\mathbf{x}_0 + \mathbf{h}) = f(\mathbf{x}_0) + J(f)_{\mathbf{x}_0}(\mathbf{h}) + \frac{1}{2} \mathbf{h}^{tr} H_{\mathbf{x}_0}(f) \mathbf{h} + \varepsilon(\mathbf{h})\|\mathbf{h}\|^2$$

Remarks 7.3.12 If A is an $n \times n$ -matrix, then the map $B : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ defined by

$$B(\mathbf{h}, \mathbf{k}) := \mathbf{h}^{tr} A \mathbf{k}$$

is an example of a *bilinear map*, i.e. one which is linear in both components:

$$\begin{aligned} B(\mathbf{x} + \mathbf{y}, \mathbf{k}) &= B(\mathbf{x}, \mathbf{k}) + B(\mathbf{y}, \mathbf{k}) & B(\mathbf{h}, \mathbf{x} + \mathbf{y}) &= B(\mathbf{h}, \mathbf{x}) + B(\mathbf{h}, \mathbf{y}) \\ B(\alpha \mathbf{x}, \mathbf{k}) &= \alpha B(\mathbf{x}, \mathbf{k}) & B(\mathbf{h}, \alpha \mathbf{x}) &= \alpha B(\mathbf{h}, \mathbf{x}) \end{aligned}$$

In particular, the map

$$D^2 f(\mathbf{x}_0) : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R} : (\mathbf{h}, \mathbf{k}) \mapsto \mathbf{h}^{tr} H_{\mathbf{x}_0} \mathbf{k}$$

is a bilinear map, called the *second derivative* of f at \mathbf{x}_0 . It can be shown that it is, indeed, the derivative of the Df , when interpreted in the right way, but we will not pursue this further. We now have

$$f(\mathbf{x}_0 + \mathbf{h}) = f(\mathbf{x}_0) + Df(\mathbf{x}_0)(\mathbf{h}) + \frac{1}{2} D^2 f(\mathbf{x}_0)(\mathbf{h}, \mathbf{h}) + \varepsilon(\mathbf{h})\|\mathbf{h}\|^2$$

In a similar way, the third-order terms can be summarised by a *trilinear map* $D^3f(\mathbf{x}_0)$, etc. (but such multilinear maps cannot be represented by matrices). In any case, the higher-order Taylor Theorems can then be written in the following form:

$$f(\mathbf{x}_0 + \mathbf{h}) = f(\mathbf{x}_0) + Df(\mathbf{x}_0)(\mathbf{h}) + \frac{1}{2!}D^2f(\mathbf{x}_0)(\mathbf{h}, \mathbf{h}) + \cdots + \frac{1}{n!}D^n f(\mathbf{x}_0)(\mathbf{h}, \dots, \mathbf{h}) + \varepsilon(\mathbf{h})\|\mathbf{h}\|^n$$

where $\varepsilon(\mathbf{h}) \rightarrow 0$ as $\mathbf{h} \rightarrow 0$, and each $D^k f(\mathbf{x}_0)$ is a k -multilinear mapping.

□

7.4 Maxima and Minima

7.4.1 Topological Facts about Extrema

Definition 7.4.1 Suppose that $C \subseteq \mathbb{R}^n$ that $f : C \rightarrow \mathbb{R}$, and that $\mathbf{x}_0 \in C$. If $f(\mathbf{x}_0) \geq f(\mathbf{x})$ for all $\mathbf{x} \in C$, then we say that \mathbf{x}_0 is a *global maximum* of f on C . Moreover, if $f(\mathbf{x}_0) > f(\mathbf{x})$ for all $\mathbf{x} \in C$, then we say that \mathbf{x}_0 is a *strict global maximum* of f . The notion of a (strict) global minimum is defined in a similar way.

We begin with some general topological facts about maxima and minima.

Lemma 7.4.2 If $C \subseteq \mathbb{R}$ is compact, then C contains a maximum and a minimum element.

Exercise 7.4.3 Prove Lemma 7.4.2.

[Hint: Recall the Heine–Borel Theorem. Show that because C is bounded, $\sup C < \infty$, and because C is closed, $\sup C \in C$.]

□

Theorem 7.4.4 (a) Suppose that $K \subseteq \mathbb{R}^n$ is compact, and that $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is continuous. Then

$$f[K] = \{f(x) : x \in \mathbb{R}^n\} \quad \text{is compact in } \mathbb{R}$$

(b) If $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is continuous and $K \subseteq \mathbb{R}^n$ is compact, then f attains its global maximum and –minimum on K , i.e. there exist $k_{\min}, k_{\max} \in K$ such that

$$f(k_{\min}) = \inf\{f(k) : k \in K\} \quad f(k_{\max}) = \sup\{f(k) : k \in K\}$$

Proof: (a) Suppose that $\mathcal{U} = \{U_i : i \in I\}$ is an open cover of $f[K]$ in \mathbb{R} . We must show that there exist finitely many $i_1, \dots, i_n \in I$ so that $f[K] \subseteq U_{i_1} \cup \cdots \cup U_{i_n}$. As the sets U_i are open and f is continuous, the sets $V_i := f^{-1}[U_i]$ are open in \mathbb{R}^n — cf. Proposition ???. Now

$$\begin{aligned} x &\in K \\ \Rightarrow f(x) &\in f[K] \\ \Rightarrow f(x) &\in \bigcup_{i \in I} U_i \\ \Rightarrow x &\in f^{-1}\left[\bigcup_{i \in I} U_i\right] = \bigcup_{i \in I} V_i \end{aligned}$$

Thus $K \subseteq \bigcup_{i \in I} V_i$, i.e. $\{V_i : i \in I\}$ is an open cover of K .

As K is compact, there are finitely many $i_1, \dots, i_n \in I$ so that $K \subseteq V_{i_1} \cup \dots \cup V_{i_n} = f^{-1}[U_{i_1} \cup \dots \cup U_{i_n}]$. It is now easy to see that $f[K] \subseteq U_{i_1} \cup \dots \cup U_{i_n}$.

(b) By Lemma 7.4.2, $f[K]$ has a minimum and a maximum.

◊

With these general topological facts out of the way, we can begin to study the maxima and minima of multivariate functions using calculus.

7.4.2 Maxima and Minima via Calculus

Recall two familiar facts from elementary calculus:

Fact I. $f : \mathbb{R} \rightarrow \mathbb{R}$ has a (local) extremum (i.e. maximum or minimum) at x_0 , and if f is differentiable at x_0 , then $f'(x_0) = 0$.

Fact II. We can say more if f is twice-differentiable at x_0 : If $f''(x_0) < 0$, then x_0 is a local maximum, and if $f''(x_0) > 0$, it is a local minimum.

We want to generalize these results to functions $f : \mathbb{R}^n \rightarrow \mathbb{R}$.

A generalization of the first fact is rather easy to obtain, once we've made some definitions:

Definition 7.4.5 (a) Suppose that $U \subseteq \mathbb{R}^n$ is open, that $f : U \rightarrow \mathbb{R}$, and that $\mathbf{x}_0 \in U$. If there is a neighbourhood V of \mathbf{x}_0 so that $V \subseteq U$ and

$$f(\mathbf{x}_0) \geq f(\mathbf{x}) \quad \text{for all } \mathbf{x} \in V$$

then we say that f has a *local maximum* at \mathbf{x}_0 .

The notion of a *local minimum* is defined in a similar way.

(b) \mathbf{x}_0 is called an *extreme point* of f iff it is either a local maximum or -minimum.

(c) \mathbf{x}_0 is called a *critical point* of f if f is differentiable at \mathbf{x}_0 and $Df(\mathbf{x}_0) = 0$.

Thus f has a local maximum at \mathbf{x}_0 iff there is a neighbourhood V of \mathbf{x}_0 such that

$$f(\mathbf{x}_0) = \max_{\mathbf{x} \in V} f(\mathbf{x})$$

iff \mathbf{x}_0 is a global maximum of the restriction $f|_V$ of f to V .

Here is the promised generalization of Fact I.

Theorem 7.4.6 Suppose that $U \subseteq \mathbb{R}^n$ is an open set, and that $f : U \rightarrow \mathbb{R}$ is differentiable. If \mathbf{x}_0 is an extreme point of f , then \mathbf{x}_0 is a critical point, i.e. $Df(\mathbf{x}_0) = 0$.

Proof: Suppose that $Df(\mathbf{x}_0) \neq 0$. Then there is \mathbf{v} so that $c := Df(\mathbf{x}_0)(\mathbf{v}) > 0$. As U is an open neighbourhood of \mathbf{x}_0 , there is $r > 0$ such that the ball $B(\mathbf{x}_0, r\|\mathbf{v}\|)$ is a subset of U . The map

$$g : (-r, r) \rightarrow \mathbb{R} : t \mapsto f(\mathbf{x}_0 + t\mathbf{v})$$

is now well-defined, and differentiable: Indeed, for $-r < t < r$, we have

$$g'(t) = \lim_{h \rightarrow 0} \frac{g(t+h) - g(t)}{h} = \lim_{h \rightarrow 0} \frac{f(\mathbf{x}_0 + t\mathbf{v} + h\mathbf{v}) - f(\mathbf{x}_0 + t\mathbf{v})}{h} = D_{\mathbf{v}}f(\mathbf{x}_0 + t\mathbf{v}) = Df(\mathbf{x}_0 + t\mathbf{v})(\mathbf{v})$$

which exists, because $\mathbf{x}_0 + t\mathbf{v} \in U$ when $|t| < r$, and because f is differentiable on U . In particular, $g'(0) = D_{\mathbf{v}}f(\mathbf{x}_0) = Df(\mathbf{x}_0)(\mathbf{v}) = c > 0$. By single-variable calculus, 0 is neither a local maximum or local minimum of g on $(-r, r)$. It follows that, for any $\delta > 0$, there are $t_1, t_2 \in (-\delta, \delta)$ so that

$$g(t_1) < g(0) < g(t_2)$$

from which we see that

$$f(\mathbf{x}_0 + t_1\mathbf{v}) < f(\mathbf{x}_0) < f(\mathbf{x}_0 + t_2\mathbf{v})$$

Thus there are points $\mathbf{x}_i = \mathbf{x}_0 + t_i\mathbf{v}$ ($i = 1, 2$) lying arbitrarily close to \mathbf{x}_0 — within $\delta\|\mathbf{v}\|$ for arbitrarily small δ — for which $f(\mathbf{x}_1) > f(\mathbf{x}_0)$ and $f(\mathbf{x}_2) < f(\mathbf{x}_0)$. Thus \mathbf{x}_0 is neither a local maximum or minimum of f .

—

Remarks 7.4.7 For functions $f; \mathbb{R}^2 \rightarrow \mathbb{R}$, the above result is easy to interpret graphically: We saw earlier that the tangent plane to the surface $z = f(x, y)$ at the point (x_0, y_0) has equation

$$z = f(x_0, y_0) + \partial_1 f(x_0, y_0)(x - x_0) + \partial_2 f(x_0, y_0)(y - y_0)$$

(cf. Remarks 7.2.35.) If (x_0, y_0) is a critical point of f , then $Df(x_0, y_0) = 0$ and so both partial derivatives $\partial_1 f(x_0, y_0), \partial_2 f(x_0, y_0)$ are zero. Hence the equation of the tangent plane is $z = f(x_0, y_0) = \text{constant}$, i.e. the tangent plane is horizontal. The converse is also easily seen to be true: If the tangent plane to the graph of f is horizontal at (x_0, y_0) , then (x_0, y_0) is a critical point.

□

Having successfully generalized the first fact about maxima and minima from elementary calculus, we now proceed with the generalization of Fact II.: If x_0 is a critical point of f and $f''(x_0) > 0$ (resp. < 0), then x_0 is a local minimum (resp. local maximum). We immediately encounter two problems:

- How do we interpret the second derivative f'' if $f: \mathbb{R}^n \rightarrow \mathbb{R}$?

We have already made some progress towards this when we introduced the *Hessian* matrix in connection with the second-order Taylor expansion of a function f , i.e. we may interpret the second derivative of f at \mathbf{x}_0 as the $n \times n$ -matrix

$$H_{\mathbf{x}_0}(f) = (f_{,ij}(\mathbf{x}_0))_{ij}$$

The second problem is the following:

- Now that we have an interpretation of the second derivative, how do we meaningfully generalize

$$f''(\mathbf{x}_0) < 0 \quad \text{or} \quad f''(\mathbf{x}_0) > 0 \quad ?$$

The next example paves the way:

Example 7.4.8 Suppose that $f: \mathbb{R} \rightarrow \mathbb{R}$ is twice differentiable, that $f'(x_0) = 0$ and that $f''(x_0) > 0$. The second-order Taylor expansion of f about x_0 is then

$$f(x_0 + h) = f(x_0) + 0 \cdot h + \frac{1}{2}f''(x_0)h^2 + \varepsilon(h)h^2$$

where, as usual, $\varepsilon(h) \rightarrow 0$ as $h \rightarrow 0$. In particular, there is $\delta > 0$ such that $|\varepsilon(h)| \leq \frac{1}{2}|f''(x_0)|$ whenever $|h| < \delta$. Since $f''(x_0)$ is positive, we see that $-\frac{1}{2}f''(x_0) \leq \varepsilon(h) \leq \frac{1}{2}f''(x_0)$, and thus that $f''(x_0) \geq \frac{1}{2}f''(x_0) + \varepsilon(h) \geq 0$. Thus, for sufficiently small h , we have

$$f(x_0 + h) - f(x_0) = (\tfrac{1}{2}f''(x_0) + \varepsilon(h))h^2 \geq 0 \quad \text{i.e.} \quad f(x_0 + h) \geq f(x_0)$$

and thus x_0 is a local minimum of f .

Now suppose we attempt to copy this argument. Suppose that $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is differentiable, and that \mathbf{x}_0 is a critical point of f . The second-order Taylor expansion of f about \mathbf{x}_0 is given by:

$$f(\mathbf{x}_0 + \mathbf{h}) = f(\mathbf{x}_0) + \tfrac{1}{2}\mathbf{h}^{tr}H_{\mathbf{x}_0}(f)\mathbf{h} + \varepsilon(\mathbf{h})\|\mathbf{h}\|^2$$

where $\varepsilon(\mathbf{h}) \rightarrow 0$ as $\mathbf{h} \rightarrow 0$.

Now because of this, we ought to have $|\varepsilon(\mathbf{h})| \cdot \|\mathbf{h}\|^2 \leq \tfrac{1}{2}|\mathbf{h}^{tr}H_{\mathbf{x}_0}(f)\mathbf{h}|$, for sufficiently \mathbf{h} sufficiently close to 0. (We'll explain this later, in the proof of Theorem 7.4.12.) Thus, for \mathbf{x}_0 to be a local minimum, it is enough to insist that

$$\mathbf{h}^{tr}H_{\mathbf{x}_0}(f)\mathbf{h} > 0$$

holds for all \mathbf{h} sufficiently close to 0. However, it is rather obvious that it holds for all sufficiently small \mathbf{h} , it will hold for *all* \mathbf{h} : By choosing λ sufficiently small, we can make $\lambda\mathbf{h}$ as close to 0 as we like, and

$$(\lambda\mathbf{h})^{tr}H_{\mathbf{x}_0}(\lambda\mathbf{h}) = \lambda^2(\mathbf{h}^{tr}H_{\mathbf{x}_0}(f)\mathbf{h}) \quad \text{has the same sign as} \quad \mathbf{h}^{tr}H_{\mathbf{x}_0}(f)\mathbf{h}$$

□

Definition 7.4.9 An $n \times n$ -matrix A is said to be *positive semi-definite* if and only if $\mathbf{h}^{tr}A\mathbf{h} \geq 0$ for all $\mathbf{h} \in \mathbb{R}^n$. A is said to be *positive definite* if and only if $\mathbf{h}^{tr}A\mathbf{h} > 0$ for all $\mathbf{h} \neq 0$ in \mathbb{R}^n .

The concept of negative definite and negative semi-definite matrices are defined in a similar way, by reversing the inequality signs.

Example 7.4.10 In statistics, symmetric positive (semi-)definite matrices occur naturally: Every *covariance matrix* is positive semi-definite. To see this, suppose that X_1, \dots, X_n are random variables, and that Σ is their covariance matrix, i.e.

$$\Sigma_{ij} = \text{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)]$$

Note that the covariance matrix of the random variables X_1, \dots, X_n is the same as the covariance matrix of the centered random variables $X_1 - \mathbb{E}X_1, \dots, X_n - \mathbb{E}X_n$. We may therefore assume without loss of generality that X_1, \dots, X_n are centered (i.e. that $\mathbb{E}X_i = 0$ for $i = 1, \dots, n$), and thus that $\Sigma_{ij} = \mathbb{E}[X_i X_j]$.

Now observe that if $\mathbf{a} = (a_1, \dots, a_n)^{tr} \in \mathbb{R}^n$, then $Y = \mathbf{a}^{tr}\mathbf{X} = a_1X_1 + \dots + a_nX_n$ is a centered random variable with variance

$$\text{Var}(Y) = \mathbb{E}[Y^2] = \mathbb{E}\left[\sum_{i=1}^n \sum_{j=1}^n a_i a_j X_i X_j\right] = \mathbf{a}^{tr}\Sigma\mathbf{a}$$

As variances are non-negative, we see that

$$\mathbf{a}^{tr}\Sigma\mathbf{a} \geq 0 \quad \text{for all } \mathbf{a} \in \mathbb{R}^n$$

and thus that Σ is positive semi-definite.

□

Exercise 7.4.11 Suppose that A is an $n \times n$ -matrix. Show that

$$\langle\langle x, y \rangle\rangle := \langle x, Ay \rangle$$

defines an inner product $\langle\langle \cdot, \cdot \rangle\rangle$ on \mathbb{R}^n if and only if A is symmetric and (strictly) positive definite. (Here $\langle x, y \rangle$ denotes the standard Euclidean inner product, i.e. $\langle x, y \rangle := x^{tr}y$.)

□

Keeping in mind Example 7.4.8, the following proposition cannot be a surprise:

Theorem 7.4.12 Suppose that $U \subseteq \mathbb{R}^n$ is an open set, and that $f : U \rightarrow \mathbb{R}^n$. If $Df(\mathbf{x}_0) = 0$ and $D^2f(\mathbf{x}_0) := H_{\mathbf{x}_0}(f)$ is strictly positive (resp. negative) definite, then f has a local minimum (resp. maximum) at \mathbf{x}_0 .

To prove it, we need a lemma:

Lemma 7.4.13 Suppose that A is a (strictly) positive or $-$ negative definite $n \times n$ -matrix. Then there is a $C > 0$ so that

$$C\|\mathbf{x}\|^2 \leq |\mathbf{x}^{tr} A \mathbf{x}| \quad \text{for all } \mathbf{x} \in \mathbb{R}^n$$

Proof: Suppose that A is positive definite, so that $|\mathbf{x}^{tr} A \mathbf{x}| = \mathbf{x}^{tr} A \mathbf{x}$. Define $g : \mathbb{R}^n \rightarrow \mathbb{R}$ by

$$g(\mathbf{x}) = \mathbf{x}^{tr} A \mathbf{x}$$

It is clear that g is continuous. The set $K = \{\mathbf{k} \in \mathbb{R}^n : \|\mathbf{k}\| = 1\}$ is compact in \mathbb{R}^n . By Theorem 7.4.4, g attains its minimum on K , i.e. there is $\mathbf{k}_{\min} \in K$ so that $\|\mathbf{k}_{\min}\| = 1$ and so that

$$g(\mathbf{k}_{\min}) \leq g(\mathbf{k}) \quad \text{for all } \mathbf{k} \in K$$

Since $0 \notin K$, we see that $\mathbf{k}_{\min} \neq 0$, and so that $C := g(\mathbf{k}_{\min}) > 0$ (because $H_{\mathbf{x}_0}(f)$ is strictly positive definite).

Now if $0 \neq \mathbf{x} \in \mathbb{R}^n$, then $\hat{\mathbf{x}} := \frac{\mathbf{x}}{\|\mathbf{x}\|} \in K$. So $g(\mathbf{x}) = \mathbf{x}^{tr} A \mathbf{x} = \|\mathbf{x}\|^2 (\hat{\mathbf{x}}^{tr} A \hat{\mathbf{x}}) \geq \|\mathbf{x}\|^2 C$.

+

Proof of Theorem 7.4.12: The argument provided in Example 7.4.8 is adequate. The only thing we need to verify is that $|\varepsilon(\mathbf{h})| \cdot \|\mathbf{h}\|^2 \leq \frac{1}{2} \mathbf{h}^{tr} H_{\mathbf{x}_0}(f) \mathbf{h}$, for sufficiently \mathbf{h} sufficiently close to 0.

By the preceding lemma, there is a constant $C > 0$ so that

$$\mathbf{x}^{tr} H_{\mathbf{x}_0}(f) \mathbf{x} \geq C\|\mathbf{x}\|^2 \quad \text{for all } \mathbf{x} \in \mathbb{R}^n$$

Since $\varepsilon(\mathbf{h}) \rightarrow 0$ as $\mathbf{h} \rightarrow 0$, we can find a $\delta > 0$ so that $|\varepsilon(\mathbf{h})| < \frac{1}{2}C$ whenever $\|\mathbf{h}\| < \delta$. It follows that

$$|\varepsilon(\mathbf{h})| \cdot \|\mathbf{h}\|^2 \leq C\|\mathbf{h}\|^2 \leq \frac{1}{2} \mathbf{h}^{tr} H_{\mathbf{x}_0}(f) \mathbf{h} \quad \text{whenever} \quad \|\mathbf{h}\| < \delta$$

+

We have now succeeded in generalizing Fact II: $f''(\mathbf{x}_0)$ becomes $H_{\mathbf{x}_0}(f)$, and

$$f''(\mathbf{x}_0) > 0 \quad \text{becomes} \quad H_{\mathbf{x}_0}(f) \text{ is positive definite}$$

It remains to present some results which allow one to recognize whether or not a symmetric matrix A is positive- or negative definite. Recall the following results from linear algebra:

Theorem 7.4.14 (Spectral Theorem) An $n \times n$ -matrix A is symmetric if and only if \mathbb{R}^n has an orthonormal basis of eigenvectors of A . In particular, A is diagonalizable.

Corollary 7.4.15 *An $n \times n$ -matrix A is symmetric and positive definite (resp. semi-definite) if and only if all the eigenvalues of A are strictly positive (resp. non-negative). Similarly, an $n \times n$ -matrix A is symmetric and negative definite (resp. semi-definite) if and only if all the eigenvalues of A are strictly negative (resp. non-positive).*

Proof: Since A is symmetric, there exists an orthonormal basis $\{\mathbf{v}_i : i = 1, \dots, n\}$ of eigenvectors of A . Let λ_i be the corresponding eigenvalues, i.e. $A\mathbf{v}_i = \lambda_i\mathbf{v}_i$. Then

$$\mathbf{v}_i^{tr} A \mathbf{v}_j = \lambda_i \mathbf{v}_i^{tr} \mathbf{v}_j = \lambda_i \delta_{ij}$$

It follows that $\mathbf{a} \in \mathbb{R}^n$, then $\mathbf{a} = \sum_{i=1}^n \alpha^i \mathbf{v}_i$ for some $\alpha^i \in \mathbb{R}$, and so

$$\mathbf{a}^{tr} A \mathbf{a} = \sum_{i,j=1}^n \alpha^i \alpha^j (\mathbf{v}_i^{tr} A \mathbf{v}_j) = \sum_{i,j=1}^n \alpha^i \alpha^j \lambda_j \delta_{ij} = \sum_{i=1}^n (\alpha^i)^2 \lambda_i$$

In particular, if A is positive definite, then $0 < \mathbf{v}_i^{tr} A \mathbf{v}_i = \lambda_i$, i.e. all the eigenvalues of A are strictly positive. Similarly, if A is positive semi-definite, then $0 \leq \mathbf{v}_i^{tr} A \mathbf{v}_i = \lambda_i$, so that the eigenvalues are non-negative.

Conversely, suppose that all the eigenvalues λ_i are strictly positive. If $\mathbf{a} := \sum_{i=1}^n \alpha^i \mathbf{v}_i \neq 0$, then $\alpha^i \neq 0$ for some i , and $\mathbf{a}^{tr} A \mathbf{a} = \sum_{i=1}^n (\alpha^i)^2 \lambda_i > 0$. Hence A is positive definite. Similarly, if all the eigenvalues are merely non-negative, then $\mathbf{a}^{tr} A \mathbf{a} \geq 0$, and so A is positive semi-definite.

The proof for the negative (semi-)definite case is similar.

◄

Here is a slight strengthening of Theorem 7.4.12:

Theorem 7.4.16 *Suppose that $Df(\mathbf{x}_0) = 0$ and that $H_{\mathbf{x}_0}(f)$ is non-singular. Then f has a local minimum (resp. maximum) at \mathbf{x}_0 if and only if $H_{\mathbf{x}_0}(f)$ is strictly positive (resp. negative) definite.*

Exercise 7.4.17 Prove Theorem 7.4.16.

[Hints: Note that the (\Rightarrow) -direction of the Theorem is just Theorem 7.4.12. We therefore need only prove the (\Leftarrow) -direction. Suppose therefore that $H := H_{\mathbf{x}_0}(f)$ is non-singular, and that f has a local minimum at \mathbf{x}_0 . Using a relationship between the determinant of a matrix and its eigenvalues, explain why all the eigenvalues of H are non-zero. Now suppose that one of the eigenvalues λ_i (belonging to an eigenvector \mathbf{v}_i) is strictly negative. To show that \mathbf{x}_0 is not a local minimum of f , it suffices to show that $f(\mathbf{x}_0 + t\mathbf{v}_i) < f(\mathbf{x}_0)$ for all sufficiently small t . Apply Lemma 7.4.13 to conclude this result — the 1×1 -“matrix” λ_i is negative definite.]

□

Remarks 7.4.18 If f has a Hessian $H_{\mathbf{x}_0}(f)$ at a critical point \mathbf{x}_0 which has both positive and negative eigenvalues, then f may have neither a maximum nor a minimum at \mathbf{x}_0 . In that case, \mathbf{x}_0 is called a saddle point.

□

7.4.3 Linear Least Squares

It is well-known that a parabola is completely determined by specifying three of its points, i.e. if one knows three points $(t_1, y_1), (t_2, y_2), (t_3, y_3)$ lying on a parabola $y = at^2 + bt + c$, then one can find the parabola: Just solve

$$\begin{pmatrix} t_1^2 & t_1 & 1 \\ t_2^2 & t_2 & 1 \\ t_3^2 & t_3 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

for a, b, c . Assuming none of the t_i 's are the same, the 3×3 -*Vandermonde* matrix is invertible, so the solution is unique.

If one observed a parabola in nature, however, e.g. the path of a comet, it would not be a good idea to determine the parabola by measuring just three points on its path, because of measurement error. It is better to observe the comet at a large number of times, and then to obtain the a, b, c which *best* fit the data. (Gauss developed the method of least squares precisely for the problem of determining orbits of celestial bodies.)

In general, given functions $\phi_1(t), \dots, \phi_n(t)$ and observations $(t_1, y_1), \dots, (t_m, y_m)$, we may try to find coefficients x_1, \dots, x_n which *best* fit the data, i.e. so that

$$y_j \approx \sum_{i=1}^n x_i \phi_i(t_j) \quad \text{for } j = 1, \dots, m$$

For the problem described above, we have $\phi_i(t) := t^i$ ($i = 0, 1, 2$), but we can take more general functions.

The word *best*, of course, needs further elucidation: For each $\mathbf{x} := (x_1, \dots, x_n)$ we obtain numbers $y_j^*(\mathbf{x}) := \sum_{i=1}^n x_i \phi_i(t_j)$. We want that $y_j \approx y_j^*(\mathbf{x})$ for all $j = 1, \dots, m$, i.e. we want to determine that n -tuple \mathbf{x} for which the combined error is as small as possible. We therefore want the *distance* between the vectors $\mathbf{y} := (y_1, \dots, y_m)$ and $y^*(\mathbf{x}) := (y_1^*, \dots, y_m^*)$ to be as small as possible, i.e. we want to find that value \mathbf{x} for which

$$\|\mathbf{y} - y^*(\mathbf{x})\|$$

is a minimum, where $\|\cdot\|$ denotes the usual Euclidean norm on \mathbb{R}^m . Because of the structure of this norm (square root of sum of squares) it is slightly more convenient to minimize $\|\mathbf{y} - y^*(\mathbf{x})\|^2$ instead, though this will of course amount to the same thing (as $x \mapsto x^2$ is strictly increasing on the positive real line). Also note that

$$y^*(\mathbf{x}) = \begin{pmatrix} \phi_1(t_1) & \dots & \phi_n(t_1) \\ \vdots & \dots & \vdots \\ \phi_1(t_m) & \dots & \phi_n(t_m) \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} =: A\mathbf{x}$$

We have here a function $f : \mathbb{R}^n \rightarrow \mathbb{R} : \mathbf{x} \mapsto \|\mathbf{y} - A\mathbf{x}\|^2$ whose minimum we want to find. Now the Euclidean norm is determined by the usual inner product, $\|\mathbf{z}\|^2 = \langle \mathbf{z}, \mathbf{z} \rangle$. Thus

$$f(\mathbf{x}) = \langle \mathbf{y}, \mathbf{y} \rangle - 2\langle \mathbf{y}, A\mathbf{x} \rangle + \langle A\mathbf{x}, A\mathbf{x} \rangle$$

To determine the minimum, we first solve $Df(\mathbf{x}) = 0$, using Leibniz' Rule:

$$Df(\mathbf{x}) = -2\langle \mathbf{y}, DA(\mathbf{x}) \rangle + 2\langle DA(\mathbf{x}), DA(\mathbf{x}) \rangle$$

which yields

$$\langle DA(\mathbf{x}), A\mathbf{x} - \mathbf{y} \rangle = 0 \quad \text{i.e.} \quad \langle DA(\mathbf{x})(\mathbf{h}), A\mathbf{x} - \mathbf{y} \rangle = 0 \text{ for all } \mathbf{h} \in \mathbb{R}^n$$

Now $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is clearly linear, so $DA(\mathbf{x}) = A$, for all \mathbf{x} . Thus

$$\langle A\mathbf{h}, A\mathbf{x} - \mathbf{y} \rangle = 0 \quad \text{for all } \mathbf{h} \in \mathbb{R}^n$$

Noting that $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^{tr} \mathbf{b}$, we see that $\mathbf{h}^{tr} A^{tr} (A\mathbf{x} - \mathbf{y}) = 0$ for all \mathbf{h} , and thus that $A^{tr} (A\mathbf{x} - \mathbf{y}) = 0$ which yields the *normal equations*

$$A^{tr} A \mathbf{x} = A^{tr} \mathbf{y}$$

which can (hopefully) be solved for \mathbf{x} , as $A^{tr} A$ is a square symmetric matrix.

A simple calculation (left as an exercise) shows that

$$H_{\mathbf{x}}(f) = A^{tr} A \quad \text{for all } \mathbf{x}$$

Hence if $A^{tr} A$ is positive definite, then the solution of $Df(\mathbf{x})$ must yield a minimum of f . It is easy to see that $A^{tr} A$ is always at least positive semi-definite, as $\mathbf{x}^{tr} A^{tr} A \mathbf{x} = \langle A\mathbf{x}, A\mathbf{x} \rangle = \|A\mathbf{x}\|^2 \geq 0$ for all \mathbf{x} . Clearly if $A\mathbf{x} \neq 0$ whenever $\mathbf{x} \neq 0$, then $A^{tr} A$ is strictly positive definite, and this will occur precisely when A is one-to-one, i.e. when the column vectors of A are linearly independent.

Now let's go back a bit, to see what is going on geometrically: We want to find an \mathbf{x} for which the quantity $\|\mathbf{y} - A\mathbf{x}\|$ is a minimum. Let V be the range (or image) of A , i.e. $V := \{\mathbf{z} \in \mathbb{R}^m : \exists \mathbf{x} \in \mathbb{R}^n (\mathbf{z} = A\mathbf{x})\}$. Clearly V is a linear subspace of \mathbb{R}^m , and is the space spanned by the column vectors of the matrix A . To say that $\|\mathbf{y} - A\mathbf{x}\|$ is a minimum is to say that $A\mathbf{x}$ is the point in V which lies closest to \mathbf{y} . Thus we are seeking the *orthogonal projection* of \mathbf{y} onto the subspace V .

It will become clear that all we need to solve this *best approximation* problem is the existence of orthogonal projection, and this exists in any Hilbert space. These ideas will be encountered again when we look at Fourier analysis, and also when we discuss conditional expectation.

Chapter 8

Products and Independence

8.1 The Monotone Class Theorem

We have already seen that σ -algebras can be quite complicated to deal with, and that in many cases, it is easier to work with simpler classes of sets. We therefore “broke up” the notion of σ -algebra into two parts, that of π -system and λ -system. For completeness, we will recall all the definitions and results below, but what you need to know is the following:

- A system \mathcal{A} of sets is a σ -algebra iff it is *both* a π -system *and* a λ -system.
- The notion of λ -system meshes very nicely with the continuity properties of measure.
- Therefore to prove that something is true for all events of a σ -algebra it often suffices to show that it is true for the events in a π -system that generates the σ -algebra. The events in such a π -system can often be very simple (e.g. they may be just intervals, or “rectangles.”)

Definition 8.1.1 Let \mathcal{C} be a collection of subsets of Ω

- (a) \mathcal{C} is called a π -system if it is closed under finite intersections.
- (b) \mathcal{C} is called a λ -system if
 - (i) $\Omega \in \mathcal{C}$;
 - (ii) $A, B \in \mathcal{C}$ and $A \subseteq B$ implies $B - A \in \mathcal{C}$;
 - (iii) If $A_1, A_2, \dots \in \mathcal{C}$ and $A_n \uparrow A$, then $A \in \mathcal{C}$.
- (c) We denote by $\pi(\mathcal{C})$ and $\lambda(\mathcal{C})$ the π -, respectively, λ -system *generated* by \mathcal{C} , i.e. the smallest π -, respectively, λ -system on Ω which contains \mathcal{C} .

[Why do $\pi(\mathcal{C})$, $\lambda(\mathcal{C})$ always exist?]

Proposition 8.1.2 A family \mathcal{C} of subsets of Ω is a σ -algebra iff it is both a π -system and a λ -system.

Proof: It is clear that a σ -algebra is also a π -system and a λ -system.

Conversely, suppose \mathcal{C} is both a π - and a λ -system. Then \mathcal{C} is closed under complementation, by (i), (ii) of Defn. 8.1.1(b). De Morgan's Laws applied to Defn. 8.1.1(a) show that \mathcal{C} is closed under *finite* unions. Finally, given $A_1, A_2, \dots \in \mathcal{C}$, let $A = \bigcup_n A_n$. Define

$$B_n = \bigcup_{m \leq n} A_m$$

Then each $B_n \in \mathcal{C}$, and $B_n \uparrow A$. Hence $A \in \mathcal{C}$, by Defn. 8.1.1(b)(iii), and thus \mathcal{C} is closed under countable unions.

◄

The following technical result often allows us to work with “easy” π -systems, instead of the “difficult” σ -algebras:

Theorem 8.1.3 (Dynkin's Lemma, Monotone Class Theorem)

(a) If \mathcal{C} is a π -system on Ω , then

$$\lambda(\mathcal{C}) = \sigma(\mathcal{C})$$

(b) Suppose that \mathcal{C} is a π -system and that \mathcal{D} is a λ -system (both on a set Ω), and also that $\mathcal{C} \subseteq \mathcal{D}$. Then $\sigma(\mathcal{C}) \subseteq \mathcal{D}$.

Proof: (a) Let $\mathcal{D} = \lambda(\mathcal{C})$. By Propn. 8.1.2, it suffices to show that \mathcal{D} is a π -system. We do this in two steps.

STEP I: Fix $C \in \mathcal{C}$, and define

$$\mathcal{D}_C = \{A \in \mathcal{D} : A \cap C \in \mathcal{D}\}$$

Then $\mathcal{C} \subseteq \mathcal{D}_C \subseteq \mathcal{D}$ (because \mathcal{C} is a π -system). We now show that $\mathcal{D}_C = \mathcal{D}$. To that end, it suffices to show that \mathcal{D}_C is a λ -system (because then \mathcal{D}_C is a λ -system containing \mathcal{C} , and \mathcal{D} is the *smallest* such). We therefore verify (i)-(iii) of Defn. 8.1.1:

(i) is obvious.

If $A, B \in \mathcal{D}_C$ and $A \subseteq B$, then $(B - A) \cap C = (B \cap C) - (A \cap C)$. But $B \cap C, A \cap C \in \mathcal{D}$ by definition of \mathcal{D}_C , and thus $(B - A) \cap C \in \mathcal{D}$, because \mathcal{D} is a λ -system. Thus $(B - A) \in \mathcal{D}_C$. Finally, if $A_1, A_2, \dots \in \mathcal{D}_C$ and $A_n \uparrow A$, then $A_1 \cap C, A_2 \cap C, \dots \in \mathcal{D}$ and $(A_n \cap C) \uparrow A \cap C$. Hence $A \cap C \in \mathcal{D}$, and so $A \in \mathcal{D}_C$.

We now know that $\mathcal{D}_C = \mathcal{D}$ for *every* $C \in \mathcal{C}$.

STEP II: Now, fix any $D \in \mathcal{D}$, and define

$$\mathcal{D}^D = \{A \in \mathcal{D} : A \cap D \in \mathcal{D}\}$$

First note that if $C \in \mathcal{C}$, then $\mathcal{D}_C = \mathcal{D}$, so $D \in \mathcal{D}_C$. It follows that $D \cap C \in \mathcal{D}$, and thus that $C \in \mathcal{D}^D$, for every $C \in \mathcal{C}$. Thus $\mathcal{C} \subseteq \mathcal{D}^D$, for all $D \in \mathcal{D}$.

It follows as above that \mathcal{D}^D is a λ -system, and thus that $\mathcal{D}^D = \mathcal{D}$, for all $D \in \mathcal{D}$.

In particular, if $A, B \in \mathcal{D}$, then $A \in \mathcal{D}^B$, and so $A \cap B \in \mathcal{D}$. This shows that \mathcal{D} is a π -system, and thus a σ -algebra.

(b) follows directly from (a). (Why?)

⊢

An important corollary of the preceding theorem is the following: Two probability measures which agree on a π -system agree also on the σ -algebra generated by that π -system.

Corollary 8.1.4 *Suppose that μ, ν are probability measures on a space (Ω, \mathcal{F}) , and that \mathcal{A} is a π -system which generates \mathcal{F} . If μ, ν agree on \mathcal{A} , then they agree on \mathcal{F} , i.e. $\mu = \nu$,*

Proof: (Outline) Use the continuity of measure to show that the set $\mathcal{D} := \{F \in \mathcal{F} : \mu F = \nu F\}$ is a λ -system which contains \mathcal{A} . Then deduce that $\mathcal{D} \subseteq \sigma(\mathcal{A}) = \mathcal{F}$.

⊢

The full power of the preceding technical results will now become apparent:

Theorem 8.1.5 (Monotone Class Theorem)
Let \mathcal{H} be a set of bounded functions from a set Ω into \mathbb{R} satisfying the following conditions:

- (i) *\mathcal{H} is a vector space.*
- (ii) *The constant function 1 belongs to \mathcal{H} .*
- (iii) *Given any sequence h_n of non-negative elements of \mathcal{H} such that $h_n \uparrow h$, if h is bounded, then $h \in \mathcal{H}$.*

Let \mathcal{A} be a π -system on Ω with the property that $I_A \in \mathcal{H}$ for every $A \in \mathcal{A}$.
Then every bounded $\sigma(\mathcal{A})$ -measurable function belongs to \mathcal{H} .

Proof: Let $\mathcal{D} = \{F \subseteq \Omega : I_F \in \mathcal{H}\}$. It is not hard to show that \mathcal{D} is a λ -system. By Dynkin's Lemma (Thm 8.1.3), $\mathcal{D} \supseteq \sigma(\mathcal{A})$.

Let h be a non-negative, bounded $\sigma(\mathcal{A})$ -measurable function, with upper bound K , i.e.

$$0 \leq h(\omega) \leq K \quad \text{for all } \omega \in \Omega$$

Let $h_n, n \in \mathbb{N}$ be a sequence of simple $\sigma(\mathcal{A})$ -measurable functions such that $h_n \uparrow h$. (Recall how this is done: If we define $h_n(\omega) := 2^{-n}[2^n h] := \sum_{k=1}^{K2^n} \frac{k-1}{2^n} I_{A(n,k)}(\omega)$ where $[x]$ denotes the largest integer $\leq x$ and $A(n,k) := \{\omega \in \Omega : \frac{k-1}{2^n} \leq h(\omega) < \frac{k}{2^n}\}$, then the h_n are simple functions with $h_n \uparrow h$. Since h is $\sigma(\mathcal{A})$ -measurable, each $A(n,k) \in \mathcal{D}$, i.e. $I_{A(n,k)} \in \mathcal{H}$.) Because \mathcal{H} is a vector space, we now see that $h_n \in \mathcal{H}$ for each $n \in \mathbb{N}$. Thus $h \in \mathcal{H}$ as well.

We have now shown that every non-negative bounded $\sigma(\mathcal{A})$ -measurable function belongs to \mathcal{H} . The same result can be obtained for arbitrary bounded h by splitting into positive and negative parts: $h = h^+ - h^-$.

⊢

The Monotone Class Theorem can be used in the same way as the “standard machine”, and is often more powerful. (Indeed, there are even more powerful monotone class theorems available, but they are not cheap, and we will not need them.)

8.2 Products

8.2.1 Introduction

Example 8.2.1 (a) Denote by μA the *area* of a subset A of \mathbb{R}^2 . We know how to define μ on *rectangles*, i.e. sets of the form $A = B_1 \times B_2$, where B_1, B_2 are intervals in \mathbb{R} : Indeed

$$\mu A = \lambda B_1 \times \lambda B_2 \quad (*)$$

where λ is Lebesgue measure. So μ is to be a measure on $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ such that $\mu(B_1 \times B_2) = \lambda(B_1)\lambda(B_2)$. Of course, many sets in $\mathcal{B}(\mathbb{R}^2)$ do not have the form $B_1 \times B_2$, and we would like μ to be defined for them as well. So $(*)$ cannot serve as a definition of μ .

- (b) In probability theory, it is quite natural to consider the product of two probability spaces. Such products typically model sequences of independent experiments. For example, let $\Omega_1 = \{H, T\}$, $\mathcal{F}_1 = \mathcal{P}(\Omega_1)$ and let $\mathbb{P}_1\{H\} = \frac{1}{2} = \mathbb{P}_1\{T\}$. Then $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ models the tossing of a fair coin. Now let $\Omega_2 = \{1, 2, \dots, 6\}$, $\mathcal{F}_2 = \mathcal{P}(\Omega_2)$ and $\mathbb{P}_2\{1\} = \mathbb{P}_2\{2\} = \dots = \mathbb{P}_2\{6\} = \frac{1}{6}$. Then $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ models the rolling of a fair die. The underlying set of the probability space which models the *combined* random experiment “First toss a fair coin, and then roll a fair die” can clearly be taken to be the cartesian product $\Omega = \Omega_1 \times \Omega_2$. The natural σ -algebra will be $\mathcal{F} = \mathcal{P}(\Omega_1 \times \Omega_2)$, and it is not hard to see that this σ -algebra is generated by the π -system $\{B_1 \times B_2 : B_1 \in \mathcal{F}_1, B_2 \in \mathcal{F}_2\}$. Now the event $B_1 \times B_2 \subseteq \Omega$ consists of all those outcomes $\omega = (\omega_1, \omega_2) \in \Omega_1 \times \Omega_2$ having $\omega_1 \in B_1$ and $\omega_2 \in B_2$. Thus $B_1 \times B_2$ occurs in the combined random experiment iff B_1 and B_2 occur in each of the individual experiments.

The probability measure associated with the combined random experiment would therefore naturally satisfy

$$\mathbb{P}(B_1 \times B_2) = \mathbb{P}_1(B_1)\mathbb{P}_2(B_2) \quad (**)$$

But not every event in $\mathcal{P}(\Omega_1 \times \Omega_2)$ is of the form $B_1 \times B_2$, so $(**)$ cannot serve as a definition of \mathbb{P} . □

The aim of this section is to construct, out of two measure spaces $(S, \mathcal{S}, \mu), (T, \mathcal{T}, \nu)$ a new measure space $(S \times T, \mathcal{S} \otimes \mathcal{T}, \mu \otimes \nu)$ satisfying the following requirements:

- (i) A subset of $S \times T$ is called a *measurable rectangle* if it has the form $A \times B$, where $A \in \mathcal{S}, B \in \mathcal{T}$.
 $\mathcal{S} \otimes \mathcal{T}$ is *defined* to be the smallest σ -algebra on $S \times T$ which has all rectangles with measurable sides as members.
- (ii) For each *rectangle* $A \times B$, we require that $(\mu \otimes \nu)(A \times B) = \mu A \cdot \nu B$

Remarks 8.2.2 (a) A remark on notation: We will be working with functions of more than one variable, and may integrate with respect to just one of those variables. We therefore introduce the following notation:

$$\int f(x) \mu(dx) := \mu f =: \mu^x f(x)$$

Thus, for example, $\int f(x, y) \mu(dx)$ integrates the function $f(x, y)$ over x , keeping y fixed. The integral $\iint f(x, y) \mu(dx) \nu(dy)$ is a double integral that first integrates f w.r.t. μ over the variable x , and then integrates the function $y \mapsto \int f(x, y) \mu(dx)$ w.r.t. ν over the variable y . We may also write this as $\nu^y(\mu^x f(x, y))$.

- (b) Several times below, we will prove a result for *finite* measures, and then refer to a “standard argument” to lift the result to σ -finite measures. This is done as follows: Suppose that μ is σ -finite on (S, \mathcal{S}) , and that a result Φ has been proved to hold for finite measures. Since μ is σ -finite, there exists a sequence of measurable sets $A_n \uparrow S$ such that $\mu A_n < \infty$ for all $n \in \mathbb{N}$. The measures $\mu_n := I_{A_n} \cdot \mu$ are *finite* on (S, \mathcal{S}) , so that result Φ holds for the μ_n . By the MCT, if $f \in m\mathcal{S}^+$, then

$$\mu f = \mu(\lim_n f I_{A_n}) = \lim_n \mu_n f$$

This is often enough to show that Φ holds for μ as well. □

8.2.2 Products of Measure Spaces

Given two measurable spaces $(S, \mathcal{S}), (T, \mathcal{T})$, we can construct a σ -algebra $\mathcal{S} \otimes \mathcal{T}$ on the cartesian product $S \times T$, as follows: Define *projections* $\pi_S : S \times T \rightarrow S$, $\pi_T : S \times T \rightarrow T$ by

$$\pi_S : (s, t) \mapsto s \quad \pi_T : (s, t) \mapsto t$$

The interpretation is as follows: (s, t) denotes a sample point in a space of “combined” outcomes: i.e. $s \in S$ occurred *and* $t \in T$ occurred. Given such a combined outcome $\omega = (s, t)$, $\pi_S(\omega) = s$ measures which outcome occurred in S , and $\pi_T(\omega) = t$ measures which outcome occurred in T . Given that we know a combined outcome $\omega = (s, t)$, we should also know the component outcomes s and t . Thus the projection mappings π_S, π_T should be measurable. The product σ -algebra $\mathcal{S} \otimes \mathcal{T}$ is defined to be the smallest σ -algebra on $S \times T$ which makes these maps measurable. To recapitulate:

Definition 8.2.3 Let (S, \mathcal{S}) and (T, \mathcal{T}) be measurable spaces. Define *projections* $\pi_S : S \times T \rightarrow S$, $\pi_T : S \times T \rightarrow T$ by

$$\pi_S : (s, t) \mapsto s \quad \pi_T : (s, t) \mapsto t$$

Then define $\mathcal{S} \otimes \mathcal{T} := \sigma(\pi_S, \pi_T)$ to be the smallest σ -algebra for which both projections are measurable.

Exercise 8.2.4 Let (S, \mathcal{S}) and (T, \mathcal{T}) be measurable spaces, and let $\mathcal{R} := \{A \times B : A \in \mathcal{S}, B \in \mathcal{T}\}$ be the set of all measurable rectangles. Note that \mathcal{R} is a π -system. Show that $\mathcal{S} \otimes \mathcal{T} = \sigma(\mathcal{R})$. Hence the product σ -algebra is generated by the π -system of all measurable rectangles. [Hint: $A \times B = (A \times T) \cap (S \times B)$, and $A \times T = \pi_S^{-1}[A]$.]

□

Exercise 8.2.5 Show that $\mathcal{B}(\mathbb{R}^2) = \mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R})$.

[Hint: Using Exercise 16.3.4, it is easy to see that $\mathcal{B}(\mathbb{R}^2) \supseteq \mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R})$. For the opposite direction, show that any open set in \mathbb{R}^2 can be written as a countable union of sets of the form $U \times V$, where U, V are open intervals in \mathbb{R} .]

□

Suppose that (S, \mathcal{S}, μ) and (T, \mathcal{T}, ν) are measure spaces. We would like to construct a measure $\mu \otimes \nu$ on $(S \times T, \mathcal{S} \otimes \mathcal{T})$. One way that suggests itself is to define

$$(1) \quad (\mu \otimes \nu)B := \int \left(\int I_B(s, t) \nu(dt) \right) \mu(ds) = \mu^s(\nu^t I_B(s, t)) \quad B \in \mathcal{S} \otimes \mathcal{T}$$

Another is to define it as

$$(2) \quad (\mu \otimes \nu)B := \int \left(\int I_B(s, t) \mu(ds) \right) \nu(dt) = \nu^t(\mu^s I_B(s, t)) \quad B \in \mathcal{S} \otimes \mathcal{T}$$

Exercise 8.2.6 Check that

$$(\mu \otimes \nu)(A \times B) = \mu A \cdot \nu B \quad A \in \mathcal{S}, B \in \mathcal{T}$$

for both of the above possible definitions of $\mu \otimes \nu$.

□

We shall soon see that (i) the above definitions are both possible, and (ii) they coincide.

We first investigate the possibility of defining $\mu \otimes \nu$ in the above manner. To be able to do perform a double integral $\iint f(s, t) \nu(dt) \mu(ds)$ it is necessary that:

- (i) for each $s \in S$, the map $t \mapsto f(s, t)$ must \mathcal{T} -measurable, so that we can calculate the inner integral $\int f(s, t) \nu(dt)$;
- (ii) the map $s \mapsto F(s) := \int f(s, t) \nu(dt)$ must be \mathcal{S} -measurable, so that we can calculate the outer integral $\int F(s) \mu(ds)$.

The following lemma gives us what we need:

Lemma 8.2.7 *Suppose that (S, \mathcal{S}) and (T, \mathcal{T}) are measurable spaces, that μ is a σ -finite measure on (S, \mathcal{S}) , and that $f : S \times T \rightarrow \mathbb{R}^+$ is $\mathcal{S} \otimes \mathcal{T}$ -measurable. Then*

- (i) *For each $t \in T$, the map $s \mapsto f(s, t)$ is \mathcal{S} -measurable.*
- (ii) *The map $t \mapsto \int f(s, t) \mu(ds)$ is \mathcal{T} -measurable.*

Proof: We apply the Monotone Class Theorem (Thm. 8.1.5). First assume that μ is a finite measure, and let

$$\mathcal{H} = \{f \in m\mathcal{S} \otimes \mathcal{T} : f \text{ is bounded and satisfies (i) and (ii)}\}$$

It is easy to verify that \mathcal{H} is a vector space (we need the finiteness of μ in order to avoid expressions of the form $\infty - \infty$), and that that each $I_{A \times B} \in \mathcal{H}$, where $A \in \mathcal{S}, B \in \mathcal{T}$. By the MCT, \mathcal{H} is closed under bounded limits of increasing non-negative sequences. Moreover, the set $\mathcal{R} := \{A \times B : A \in \mathcal{S}, B \in \mathcal{T}\}$ is a π -system with the property that $I_R \in \mathcal{H}$ for every $R \in \mathcal{R}$, and thus by Thm. 8.1.5 every bounded $\mathcal{S} \otimes \mathcal{T}$ -measurable function belongs to \mathcal{H} (since $\sigma(\mathcal{R}) = \mathcal{S} \otimes \mathcal{T}$). Now each non-negative measurable function f is the limit of bounded non-negative measurable functions ($f = \lim_n f \wedge n$), and thus another application of the MCT shows that every $f \in m(\mathcal{S} \otimes \mathcal{T})^+$ satisfies (i) and (ii).

Now drop the assumption that μ is a finite measure. Because μ is σ -finite, we can choose $A_n \uparrow S$ such that $\mu A_n < \infty$. The measures $\mu_n = I_{A_n} \cdot \mu$ are finite measures, and thus each map $t \mapsto \int f(s, t) \mu_n(ds)$ is \mathcal{T} -measurable (where $f \geq 0$). Since $\int f(s, t) \mu(ds) = \lim_n \int f(s, t) \mu_n(ds)$, the MCT implies that the result holds for μ .

⊥

We now know that it is possible to define $\mu \otimes \nu$ in the ways indicated. What we don't (yet) know is that these constructions define a *measure*, and that they *coincide*.

For definiteness, we fix one of the above definitions:

Definition 8.2.8 Suppose that (S, \mathcal{S}, μ) and (T, \mathcal{T}, ν) are σ -finite measure spaces. Define a map $\mu \otimes \nu : \mathcal{S} \otimes \mathcal{T} \rightarrow \bar{\mathbb{R}}^+$ by

$$(\mu \otimes \nu)B := \iint I_B(s, t) \nu(dt) \mu(ds) = \mu^s(\nu^t I_B(s, t)) \quad B \in \mathcal{S} \otimes \mathcal{T}$$

$\mu \otimes \nu$ is called the *product measure* of μ, ν .

□

Exercise 8.2.9 Show that $\mu \otimes \nu$ defines a σ -finite measure on $(S \times T, \mathcal{S} \otimes \mathcal{T})$.

□

The next two results show that (modulo certain conditions) we can calculate the integral w.r.t. $\mu \otimes \nu$ as a double integral, and the order of integration doesn't matter:

$$\int f d(\mu \otimes \nu) = \iint f(s, t) \nu(dt) \mu(ds) = \iint f(s, t) \mu(ds) \nu(dt)$$

We first show this for non-negative measurable functions:

Theorem 8.2.10 (Tonelli)

Suppose that (S, \mathcal{S}, μ) and (T, \mathcal{T}, ν) are σ -finite measure spaces. If $f \in m(\mathcal{S} \otimes \mathcal{T})^+$, then

$$(\mu \otimes \nu)f = \mu^s(\nu^t f(s, t)) = \nu^t(\mu^s f(s, t)) \quad (*)$$

Proof: We use the Monotone Class Theorem (Thm. 8.1.5). First assume that μ, ν are finite measures. The result is obvious if $f = I_{A \times B}$, where $A \times B$ measurable rectangle, (or cf. Exercise 16.3.6). The class

$$\mathcal{H} = \{f \in m(\mathcal{S} \otimes \mathcal{T}) : f \text{ is bounded and satisfies } (*)\}$$

is easily seen to satisfy the requirements of Thm. 8.1.3, and thus implies that \mathcal{H} contains every bounded $\mathcal{S} \otimes \mathcal{T}$ -measurable function. The result for arbitrary non-negative f follows by MCT.

A standard argument lifts the result to the case where μ, ν are merely σ -finite.

◄

As a by-product, we obtain the result that our two possible definitions of $\mu \otimes \nu$ as iterated integrals coincide: If $B \in \mathcal{S} \otimes \mathcal{T}$, then I_B is a non-negative measurable function, and we may apply Tonelli's Thm.

For non-negative functions f , the integral μf always makes sense, but we may have $\mu f = \infty$. For arbitrary measurable f , we have to be more careful:

Theorem 8.2.11 (Fubini)

Suppose that (S, \mathcal{S}, μ) and (T, \mathcal{T}, ν) are σ -finite measure spaces. If $f \in \mathcal{L}^1(S \times T, \mathcal{S} \otimes \mathcal{T}, \mu \otimes \nu)$, then

$$(\mu \otimes \nu)f = \mu^s(\nu^t f(s, t)) = \nu^t(\mu^s f(s, t))$$

Here the map $t \mapsto \mu^s f(s, t)$ belongs to $\mathcal{L}^1(T, \mathcal{T}, \nu)$ for ν -a.e. $t \in T$. Similarly, the map $s \mapsto \nu^t f(s, t)$ belongs to $\mathcal{L}^1(S, \mathcal{S}, \mu)$ for μ -a.e. $s \in S$.

Proof: The result holds for $|f|$, by Tonelli's Thm., and hence $N_S = \{s \in S : \nu^t |f(s, t)| = +\infty\}$ is μ -null, and $N_T = \{t \in T : \mu^s |f(s, t)| = +\infty\}$ is ν -null. Redefine $f(s, t)$ to be zero when either $s \in N_S$ or $t \in N_T$; this won't affect the integral of f , by Thm. 6.3.9. The result follows by splitting f into positive and negative parts.

⊥

Remarks 8.2.12 (a) Fubini's Theorem allows the interchange of the order of integration, provided the integrand is integrable w.r.t the product measure. It follows from Fubini's Theorem that

$$\int \left(\int f \, d\nu \right) d\mu = \int \left(\int f \, d\mu \right) d\nu$$

provided that $f \in \mathcal{L}^1$. See Exercise 16.3.13 for what can happen if $f \notin \mathcal{L}^1$.

(b) Fubini's Theorem also easily extends to arbitrary finite products: If $(S_i, \mathcal{S}_i, \mu_i)$ are σ -finite measure spaces for $i = 1, \dots, n$, then

(i) $\mathcal{S}_1 \otimes \dots \otimes \mathcal{S}_n$ is the σ -algebra on $S_1 \times \dots \times S_n$ which is generated by the projections $\pi_i : S_1 \times \dots \times S_n \rightarrow S_i : (s_1, \dots, s_n) \mapsto s_i$. It is also generated by the family of measurable "rectangles" $\mathcal{R} = \{A_1 \times \dots \times A_n : A_i \in \mathcal{S}_i \text{ for } i = 1, \dots, n\}$.

(ii) $\mu_1 \otimes \dots \otimes \mu_n$ is the unique measure on $\mathcal{S}_1 \otimes \dots \otimes \mathcal{S}_n$ which assigns to every rectangle the measure

$$(\mu_1 \otimes \dots \otimes \mu_n)(A_1 \times \dots \times A_n) = \mu_1 A_1 \dots \mu_n A_n$$

(iii) Fubini's Theorem states that if $f : S_1 \times \dots \times S_n \rightarrow \bar{\mathbb{R}}$ is $\mu_1 \otimes \dots \otimes \mu_n$ -integrable, then

$$\int_{S_1 \times \dots \times S_n} f \, d(\mu_1 \otimes \dots \otimes \mu_n) = \int_{S_1} \left(\int_{S_2} \dots \left(\int_{S_n} f \, d\mu_n \right) \dots d\mu_2 \right) d\mu_1$$

and that any interchange of the order of integration is permissible.

□

Exercise 8.2.13 Let

$$f(x, y) = \frac{x^2 - y^2}{(x^2 + y^2)^2}$$

Show that

$$\int_0^1 \int_0^1 f(x, y) \, \lambda(dy) \, \lambda(dx) = \frac{\pi}{4} \quad \int_0^1 \int_0^1 f(x, y) \, \lambda(dx) \, \lambda(dy) = -\frac{\pi}{4}$$

What can you conclude about

$$\int_{[0,1] \times [0,1]} f \, d(\lambda \otimes \lambda)$$

□

8.3 Independence

We return now to the notion of *independence*. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, Recall that we have made the following definitions:

- Events $F_1, \dots, F_n \in \mathcal{F}$ are independent iff $\mathbb{P}(F_1 \cap \dots \cap F_n) = \prod_{k=1}^n \mathbb{P}(F_k)$.
- Sub- σ -algebras $\mathcal{G}_1, \dots, \mathcal{G}_n$ are independent iff whenever $G_k \in \mathcal{G}_k$ for $k = 1, \dots, n$ then G_1, \dots, G_n are independent events.
- A random variable X is independent of a σ -algebra \mathcal{G} iff $\sigma(X), \mathcal{G}$ are independent σ -algebras.

Other variations (e.g. the what it means for random variables X_1, \dots, X_n to be independent) should be obvious.

If two π -systems are independent, so are the σ -algebras generated by those π -systems:

Theorem 8.3.1 *Let $\{\mathcal{C}_t\}_{t \in T}$ be a collection of independent π -systems on $(\Omega, \mathcal{F}, \mathbb{P})$. Then $\{\sigma(\mathcal{C}_t)\}_{t \in T}$ is a collection of independent σ -algebras.*

Proof: We must show that if $t_1, \dots, t_n \in T$ are distinct, then $\sigma(\mathcal{C}_{t_1}), \dots, \sigma(\mathcal{C}_{t_n})$ are independent. We proceed by recursion. Fix $t_1, \dots, t_n \in T$, define $\mathcal{F}_{t_k} := \sigma(\mathcal{C}_{t_k})$, and also fix $C_{t_2} \in \mathcal{C}_{t_2}, \dots, C_{t_n} \in \mathcal{C}_{t_n}$. Let

$$\mathcal{D} := \{F \in \mathcal{F}_{t_1} : \mathbb{P}(F \cap C_{t_2} \cap \dots \cap C_{t_n}) = \mathbb{P}F \cdot \mathbb{P}C_{t_2} \cdot \dots \cdot \mathbb{P}C_{t_n}\}$$

By assumption, $\mathcal{C}_{t_1} \subseteq \mathcal{D}$. Using the continuity of measure, it is straightforward to check that \mathcal{D} is a λ -system. Thus by Thm. 8.1.3 we have $\mathcal{D} = \mathcal{F}_{t_1}$ for every selection of $C_{t_k} \in \mathcal{C}_{t_k}$ $k = 2, 3, \dots, n$, and hence the families $\mathcal{F}_{t_1}, \mathcal{C}_{t_2}, \mathcal{C}_{t_3}, \dots, \mathcal{C}_{t_n}$ are independent. Repeat: Fix $F_{t_1} \in \mathcal{F}_{t_1}$ and $C_{t_3} \in \mathcal{C}_{t_3}, \dots, C_{t_n} \in \mathcal{C}_{t_n}$. Redefine

$$\mathcal{D} := \{F \in \mathcal{F}_{t_2} : \mathbb{P}(F_{t_1} \cap F \cap C_{t_2} \cap \dots \cap C_{t_n}) = \mathbb{P}F_{t_1} \cdot \mathbb{P}F \cdot \mathbb{P}C_{t_2} \cdot \dots \cdot \mathbb{P}C_{t_n}\}$$

Again, \mathcal{D} is a λ -system containing \mathcal{C}_{t_2} , and hence by Thm. 8.1.3 $\mathcal{D} = \mathcal{F}_{t_2}$. From this it follows that $\mathcal{F}_{t_1}, \mathcal{F}_{t_2}, \mathcal{C}_{t_3}, \dots, \mathcal{C}_{t_n}$ are independent. Repeat the construction $n - 2$ more times to deduce that $\mathcal{F}_{t_1}, \dots, \mathcal{F}_{t_n}$ are independent. ◄

If you're familiar with the elementary definition of independence for random variables, you will want to know the following:

Exercise 8.3.2 Random variables X, Y are said to be independent in the elementary sense iff

$$\mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(X \leq x) \cdot \mathbb{P}(Y \leq y) \quad \text{all } x, y \in \mathbb{R}$$

Prove that random variables are independent iff they are independent in the elementary sense.

[Hint: First show that set $\mathcal{X} := \{X \leq x\} : x \in \mathbb{R}\}$ is a π -system which generates $\sigma(X)$. ◻

If $\{\mathcal{G}_t\}_{t \in T}$ is a family of sub- σ -algebras of \mathcal{F} , define

$$\bigvee_{t \in T} \mathcal{G}_t := \sigma\left(\bigcup_{t \in T} \mathcal{G}_t\right)$$

Proposition 8.3.3 *Suppose that $\{\mathcal{G}_t\}_{t \in T}$ is a family of independent σ -algebras, and that \mathcal{P} is a partition of T . For $P \in \mathcal{P}$, define $\mathcal{G}_P := \bigvee_{t \in P} \mathcal{G}_t$. Then $\{\mathcal{G}_P\}_{P \in \mathcal{P}}$ is a family of independent σ -algebras.*

Proof: For $P \in \mathcal{P}$, define \mathcal{C}_P to be the set of all finite intersections of members of $\bigcup_{t \in P} \mathcal{G}_t$. Then each \mathcal{C}_P is a π -system, and $\sigma(\mathcal{C}_P) = \mathcal{G}_P$. The independence of the \mathcal{G}_t , $t \in T$, is easily seen to imply the independence of the \mathcal{C}_P , $P \in \mathcal{P}$, and thus the independence of the \mathcal{G}_P , $P \in \mathcal{P}$ (by Propn. 8.3.1). ◄

The following result is now easy to within the measure-theoretic framework, but very difficult to prove outside it:

Theorem 8.3.4 Suppose that X_1, \dots, X_{n+m} are independent random variables, and that $\mathbb{R}^n \xrightarrow{f} \mathbb{R}$ and $\mathbb{R}^m \xrightarrow{g} \mathbb{R}$ are Borel functions. Then $Y = f(X_1, \dots, X_n)$ and $Z = g(X_{n+1}, \dots, X_{n+m})$ are independent.

Exercise 8.3.5 Prove Thm. 8.3.4.

□

Earlier in this course, we proved the following result using a “standard machine” approach:

Theorem 8.3.6 Suppose that $X, Y \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ are independent random variables. Then $XY \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ also, and

$$\mathbb{E}XY = \mathbb{E}X \cdot \mathbb{E}Y \quad \text{i.e.} \quad \text{Cov}(X, Y) = 0$$

The same result holds in the extended sense if $X, Y \geq 0$.

Actually, there is an easier proof of Propn. 8.3.4, if we adopt another point of view: If X, Y are random variables, then (X, Y) is a random vector, i.e. a map $(\Omega, \mathcal{F}, \mathbb{P}) \xrightarrow{(X,Y)} (\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$. Its distribution is a probability measure on $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ given by $\mu_{X,Y}(B) = \mathbb{P}\{(X, Y) \in B\}$, where $B \in \mathcal{B}(\mathbb{R}^2)$. If μ_X, μ_Y are the distributions of X, Y respectively, then the product measure $\mu_X \otimes \mu_Y$ is another probability measure on $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$. It turns out that X, Y are independent iff $\mu_{X,Y} = \mu_X \otimes \mu_Y$:

Theorem 8.3.7 Let X, Y be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mu_{X,Y}, \mu_X, \mu_Y$ be the distributions of the random elements $(X, Y), X$ and Y . Then X, Y are independent iff $\mu_{X,Y} = \mu_X \otimes \mu_Y$.

Proof: Suppose that X, Y are independent. If $A \times B$ is a measurable rectangle in $\mathcal{B}(\mathbb{R}^2) = \mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R})$, then

$$\mu_{X,Y}(A \times B) = \mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B) = \mu_X A \cdot \mu_Y B = (\mu_X \otimes \mu_Y)(A \times B)$$

Hence $\mu_{X,Y}, \mu_X \otimes \mu_Y$ agree on a π -system that generates $\mathcal{B}(\mathbb{R}^2)$ (the family of measurable rectangles). Since $\mu_{X,Y}$ and $\mu_X \otimes \mu_Y$ agree on a π -system that generates $\mathcal{B}(\mathbb{R}^2)$, they are equal: $\mu_{X,Y} = \mu_X \otimes \mu_Y$.

Conversely, if $\mu_{X,Y} = \mu_X \otimes \mu_Y$, then if $x, y \in \mathbb{R}$, we have

$$\mathbb{P}(X \leq x, Y \leq y) = \mu_{X,Y}((-\infty, x] \times (-\infty, y]) = \mu_X(-\infty, x] \cdot \mu_Y(-\infty, y] = \mathbb{P}(X \leq x)\mathbb{P}(Y \leq y)$$

Hence X, Y are independent (cf. Exercise 8.3.2).

◄

Exercise 8.3.8 Use the Change of Variables Thm. to show that

$$\int XY \, d\mathbb{P} = \int xy \, d\mu_{X,Y}$$

Now prove Propn. 8.3.6 once more, using Fubini's Theorem.

□

Definition 8.3.9 Let μ, ν be probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The convolution $\mu * \nu$ of μ and ν is defined to be the pushforward of the product measure $\mu \otimes \nu$ along the measurable map $+: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R} : (x, y) \mapsto x + y$.

$$\mu * \nu := (\mu \otimes \nu)(+)^{-1} \quad \text{i.e.} \quad (\mu * \nu)B := \mu \otimes \nu\{(x, y) : x + y \in B\}$$

Exercise 8.3.10 (a) Show that

$$(\mu * \nu)B = \int \nu(B - x) \mu(dx) = \int \mu(B - y) \nu(dy)$$

where $B - x = \{b - x : b \in B\}$.

- (b) Show that $*$ is a commutative associative operation on the set of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Show that δ_0 is an identity element for $*$.
- (c) Show that if X, Y are independent random variables, then $\mu_{X+Y} = \mu_X * \mu_Y$, where μ_X is the distribution of X , etc.

□

Chapter 9

The \mathcal{L}^p –Spaces and Fourier Analysis

9.1 \mathcal{L}^p Spaces

9.1.1 Integration of complex-valued functions

This paragraph concerns the integration of complex-valued functions. Recall that any complex number $z \in \mathbb{C}$ can be decomposed into a *real* part and an *imaginary* part:

$$z = a + ib = \operatorname{Re}(z) + i\operatorname{Im}(z) \quad \text{where } \operatorname{Re}(z) = a, \operatorname{Im}(z) = b$$

We can also write z in modulus–argument form: Recall that $e^{i\theta} := \cos \theta + i \sin \theta$. Then

$$z = re^{i\theta} \quad \text{where } r = \sqrt{a^2 + b^2} = |z| \text{ and } \tan \theta = \frac{b}{a}$$

The *complex conjugate* of $z = a + ib = re^{i\theta}$ is $\bar{z} := a - ib = re^{-i\theta}$.

Similarly, if f is a complex-valued function, then f can be decomposed into a real and imaginary part

$$f = \operatorname{Re} f + i\operatorname{Im} f = u + iv$$

where $u = \operatorname{Re} f$ and $v = \operatorname{Im} f$ are real-valued functions.

Definition 9.1.1 Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, and let $f : \Omega \rightarrow \mathbb{C}$ be a complex-valued function. We say that f is \mathcal{F} –measurable if and only if both the real-valued functions $\operatorname{Re} f$ and $\operatorname{Im} f$ are \mathcal{F} –measurable.

Now note the following:

- (i) If $f = u + iv$ is \mathcal{F} –measurable, then so is its modulus, the real-valued function $|f| = \sqrt{u^2 + v^2}$.
- (ii) Note also that $\bar{f} = u - iv$ is measurable if f is.
- (iii) We have

$$\max\{|u|, |v|\} = (\max\{u^2, v^2\})^{\frac{1}{2}} \leq (u^2 + v^2)^{\frac{1}{2}} \leq (u^2 + 2|u||v| + v^2)^{\frac{1}{2}} = |u| + |v|$$

and hence that

$$|u|, |v| \leq |f| \leq |u| + |v|$$

It follows that

$$\int |f| d\mu < \infty \iff \int |\operatorname{Re}(f)| d\mu < \infty \text{ and } \int |\operatorname{Im}(f)| d\mu < \infty$$

These observations ensure that the following definition makes sense:

Definition 9.1.2 Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, and let $f : \Omega \rightarrow \mathbb{C}$ be a complex-valued function. We say that f is μ -integrable if and only if $|f|$ is μ -integrable (and hence iff $\operatorname{Re}(f), \operatorname{Im}(f)$ are both μ -integrable). We then define

$$\int f d\mu := \int \operatorname{Re} f d\mu + i \int \operatorname{Im} f d\mu$$

The following properties can easily be established and are left as an exercise:

Proposition 9.1.3 (a) If f, g are μ -integrable complex-valued functions, then so is $f + g$, and $\int f + g d\mu = \int f d\mu + \int g d\mu$.

(b) If f is μ -integrable and $c \in \mathbb{C}$, then $\int cf d\mu = c \int f d\mu$.

(c) If f is μ -integrable, then so is \bar{f} and $\int \bar{f} d\mu = \overline{\int f d\mu}$

(d) $|\int f d\mu| \leq \int |f| d\mu$

Exercise 9.1.4 Prove Propn. 9.1.3.

[Hint for (d): There is $\theta \in \mathbb{R}$ such that $\int f d\mu = |\int f d\mu| e^{i\theta}$. So $|\int f d\mu| = e^{-i\theta} \int f d\mu = \operatorname{Re} \int e^{-i\theta} f d\mu = \int \operatorname{Re}(e^{-i\theta} f) d\mu \leq \int |e^{-i\theta} f| d\mu.$]

□

9.1.2 Definition of \mathcal{L}^p -spaces

Definition 9.1.5 Suppose that $(\Omega, \mathcal{F}, \mu)$ is a measure space, and that $1 \leq p < +\infty$ (where p need not be an integer).

$\mathcal{L}^p(\Omega, \mathcal{F}, \mu)$ is the set of all (real- or complex-valued) functions $f \in \mathfrak{m}\mathcal{F}$ such that

$$\int |f|^p d\mu < +\infty$$

For $f \in \mathcal{L}^p(\Omega, \mathcal{F}, \mu)$, we define the p -norm of f to be

$$\|f\|_p = \left(\int |f|^p d\mu \right)^{\frac{1}{p}}$$

When the underlying measure space is clear from context, and there is little danger of confusion, we will write \mathcal{L}^p instead of $\mathcal{L}^p(\Omega, \mathcal{F}, \mu)$.

Remarks 9.1.6 (a) Note that \mathcal{L}^1 is just the set of all μ -integrable functions (since f is integrable if and only if $|f| = |f|^1$ is integrable).

(b) Note also that $f \in \mathcal{L}^p$ if and only if $f^p \in \mathcal{L}^1$.

□

Lemma 9.1.7 *The \mathcal{L}^p spaces are vector spaces.*

□

Exercise 9.1.8 Prove Lemma 9.1.7.

[Hint: If $1 \leq p < \infty$, then $|f + g|^p \leq (|f| + |g|)^p \leq \max\{(2|f|)^p, (2|g|)^p\} \leq 2^p(|f|^p + |g|^p)$.]

□

It will become clear that the \mathcal{L}^p spaces are *almost* Banach spaces (with norms $\|\cdot\|_p$), and that \mathcal{L}^2 is *almost* a Hilbert space. For our purposes, the most important spaces are \mathcal{L}^1 and \mathcal{L}^2 , and we shall give a complete separate account of the theory for these spaces.

9.1.3 \mathcal{L}^1 and \mathcal{L}^2

Consider the \mathcal{L}^1 -norm $\|\cdot\|_1$ on $\mathcal{L}^1(\Omega, \mathcal{F}, \mu)$. We do not yet know that it is a norm. However:

- If f is μ -integrable, then $\|f\|_1 \geq 0$, and $\|0\|_1 = 0$.
- If f is μ -integrable and $c \in \mathbb{C}$, then clearly

$$\|cf\|_1 = \int |cf| d\mu = |c| \int |f| d\mu = |c| \|f\|_1$$

- The triangle inequality for complex moduli (or for absolute values in the real case) yields the triangle inequality for $\|\cdot\|_1$:

$$\|f + g\|_1 := \int |f + g| d\mu \leq \int |f| + |g| d\mu = \int |f| d\mu + \int |g| d\mu = \|f\|_1 + \|g\|_1$$

There is one more condition that must be satisfied for $\|\cdot\|_1$ to be a norm, and that is

- $\|f\|_1 = 0$ if and only if $f = 0$.

This condition *fails*, however, but it is nearly true: If $\|f\|_1 = 0$, then $\int |f| d\mu = 0$. Now as $|f| \geq 0$, we know that this implies that $|f| = 0$ μ -a.e., and thus that $f = 0$ μ -a.e. We thus have

- $\|f\|_1 = 0$ if and only if $f = 0$ μ -a.e.

Thus, provided we consider two functions $f, g \in \mathcal{L}^1(\Omega, \mathcal{F}, \mu)$ to be the same when we have merely $f = g$ μ -a.e., the function $\|\cdot\|_1$ is a norm on $\mathcal{L}^1(\Omega, \mathcal{F}, \mu)$.

Every norm has associated with it a notion of convergence. If $f_n, f \in \mathcal{L}^1(\Omega, \mathcal{F}, \mu)$ for $n \in \mathbb{N}$,

$$f_n \xrightarrow{\mathcal{L}^1} f \iff \|f_n - f\|_1 \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

We say that f_n converges to f in *mean*, or in \mathcal{L}^1 .

Exercise 9.1.9 Thus far, we have only used one notion of convergence, namely μ -almost everywhere convergence. This is pointwise convergence, possibly excluding a set of measure zero, i.e.

$$f_n \rightarrow f \text{ } \mu\text{-a.e.} \iff \mu\{\omega \in \Omega : f_n(\omega) \not\rightarrow f(\omega)\} = 0$$

This is the type of convergence used in the Monotone- and Dominated Convergence Theorems. We now have a new notion of convergence (actually, we have a new notion for every \mathcal{L}^p -norm), and the two are quite different:

(a) Let $(\Omega, \mathcal{F}, \mu) = ((0, 1], \mathcal{B}((0, 1]), \lambda)$, and define $f_n = nI_{(0, \frac{1}{n}]}$. Show that $f_n \rightarrow 0$ μ -a.e., but that $(f_n)_n$ does not converge in \mathcal{L}^1 .

(b) Let $(\Omega, \mathcal{F}, \mu) = ((0, 1], \mathcal{B}((0, 1]), \lambda)$. Enumerate subintervals of $(0, 1]$ as follows:

$$A_1 = (0, 1] \quad A_2 = (0, \tfrac{1}{2}] \quad A_3 = (\tfrac{1}{2}, 1] \quad A_4 = (0, \tfrac{1}{4}] \quad A_5 = (\tfrac{1}{4}, \tfrac{1}{2}] \quad A_6 = (\tfrac{1}{2}, \tfrac{3}{4}] \quad A_7 = (\tfrac{3}{4}, 1] \quad A_8 = (0, \tfrac{1}{8}] \quad \dots$$

so that

$$A_{2^n+k} = (\tfrac{k}{2^n}, \tfrac{k+1}{2^n}] \quad \text{for } n, k \in \mathbb{N}, \quad 0 \leq k < 2^n$$

Now define $f_n = I_{A_n}$. Show that $f_n \rightarrow 0$ in \mathcal{L}^1 , but that $(f_n)_n$ does not converge μ -a.e.

□

Remarks 9.1.10 Note that the Monotone Convergence Theorem and Lebesgue Dominated Convergence Theorems give conditions which ensure that μ -a.e. implies \mathcal{L}^1 -convergence. For example, the LDCT states that if $h_n \rightarrow h$ μ -a.e., and if there is an μ -integrable g such that $|h_n| \leq g$ μ -a.e. for all n , then $\int h_n d\mu \rightarrow \int h d\mu$. Now if suppose that $f_n \rightarrow f$ μ -a.e., and that there is an μ -integrable g such that $|f_n| \leq g$ μ -a.e. Define $h_n := |f_n - f|$, $h := 0$. Then $|h_n| \leq |f_n| + |f| \leq 2g$, and $2g$ is integrable also. Thus $\int |h_n| d\mu \rightarrow \int h d\mu$, i.e. $\|f_n - f\|_1 \rightarrow 0$, i.e. $f_n \rightarrow f$ in \mathcal{L}^1 .

□

A sequence $(f_n)_{n \in \mathbb{N}}$ is a *Cauchy sequence* in \mathcal{L}^1 if and only if $\|f_n - f_m\|_1 \rightarrow 0$ as $n, m \rightarrow \infty$ (i.e. iff $\sup_{m, n \geq N} \|f_n - f_m\|_1 \rightarrow 0$ as $N \rightarrow \infty$). Now recall that a complete normed vector space — that is, a vector space in which every Cauchy sequence converges — is called a *Banach space*. We will show that \mathcal{L}^1 is a Banach space, when equipped with the \mathcal{L}^1 -norm. First, we need a lemma:

Lemma 9.1.11 *If $\sum_{n=1}^{\infty} \int |f_n| d\mu < \infty$, then $\sum_{n=1}^{\infty} f_n$ converges absolutely μ -a.e. Moreover, the function $\sum_{n=1}^{\infty} f_n$ is integrable, and $\int \sum_{n=1}^{\infty} f_n d\mu = \sum_{n=1}^{\infty} \int f_n d\mu$.*

Proof: Define $g := \sum_{n=1}^{\infty} |f_n| = \lim_{m \rightarrow \infty} \sum_{n=1}^m |f_n|$. By the Monotone Convergence Theorem and the linearity of the integral, $\int g d\mu = \sum_{n=1}^{\infty} \int |f_n| d\mu < \infty$ and hence g is integrable. In particular, $|g| < \infty$ μ -a.e., and hence $\sum_{n=1}^{\infty} f_n$ converges absolutely μ -a.e. Now as $|\sum_{n=1}^m f_n| \leq \sum_{n=1}^m |f_n| \leq g$ for all $m \in \mathbb{N}$, and g is integrable, the Lebesgue Dominated Convergence Theorem shows that $\int \sum_{n=1}^{\infty} f_n d\mu = \sum_{n=1}^{\infty} \int f_n d\mu$

⊥

Theorem 9.1.12 (a) *Suppose that $(f_n)_n$ is a Cauchy sequence in $\mathcal{L}^1(\Omega, \mathcal{F}, \mu)$. Then there exists a subsequence $(f_{n_k})_k$ and an $f \in \mathcal{L}^1$ such that $f_{n_k} \rightarrow f$ μ -a.e. as $k \rightarrow \infty$. Furthermore, the original sequence converges in \mathcal{L}^1 : $\|f_n - f\|_1 \rightarrow 0$ as $n \rightarrow \infty$.*

(b) (Riesz–Fischer Theorem for \mathcal{L}^1) ($\mathcal{L}^1(\Omega, \mathcal{F}, \mu), \|\cdot\|_1$) is a Banach space.

Proof: Suppose that $(f_n)_n$ is a Cauchy sequence in \mathcal{L}^1 . Thus for any $\varepsilon > 0$, we may pick $N \in \mathbb{N}$ such that $\sup_{m, n \geq N} \|f_m - f_n\|_1 < \varepsilon$. We choose a strictly increasing sequence $(n_k)_k$ in \mathbb{N} as follows: First, choose n_1 to be an N that works for $\varepsilon = \frac{1}{4}$, i.e. choose n_1 so that $\sup_{m, n \geq n_1} \|f_m - f_n\|_1 < \frac{1}{4}$. Next, for $k > 1$ n_k is an N that works for $\varepsilon = \frac{1}{2^{k+1}}$, i.e. choose $n_k > n_{k-1}$ so that $\sup_{m \geq n_k} \|f_m - f_{n_k}\|_1 < \frac{1}{2^{k+1}}$.

Note that $\|f_{n_{k+1}} - f_{n_k}\|_1 < \frac{1}{2^{k+1}}$ for all $k \in \mathbb{N}$. Define

$$g_0 := 0 \quad g_k := f_{n_k} - f_{n_{k-1}} \text{ for } k > 0$$

so that

$$f_{n_k} = \sum_{i=0}^k g_i \quad \text{and} \quad \|g_i\| < \frac{1}{2^i}$$

Hence $\sum_{i=0}^{\infty} \int |g_i| d\mu < \infty$. By the preceding lemma, $\sum_{i=0}^{\infty} g_i$ converges (absolutely) μ -a.e., and $f := \sum_{i=0}^{\infty} g_i \in \mathcal{L}^1$. As $f = \sum_{i=0}^{\infty} g_i = \lim_{k \rightarrow \infty} \sum_{i=0}^k g_i = \lim_{k \rightarrow \infty} f_{n_k}$, we see that $f_{n_k} \rightarrow f$ converges μ -a.e.

To finish (a), we still need to show that $\|f_n - f\|_1 \rightarrow 0$ as $n \rightarrow \infty$. Suppose that $\varepsilon > 0$. Choose $N \in \mathbb{N}$ so that $\sup_{m, n \geq N} \|f_m - f_n\|_1 < \varepsilon$. Now fix an $n \geq N$ (but allow m to vary). Note that $\lim_k |f_{n_k} - f_n| = |f - f_n|$, and thus that $\liminf_m |f_m - f_n| \leq |f - f_n|$. It follows by Fatou's Lemma that

$$\|f - f_n\|_1 = \int |f - f_n| d\mu \leq \int \liminf_m |f_m - f_n| d\mu \leq \liminf_m \int |f_m - f_n| d\mu < \varepsilon$$

and hence that $\|f - f_n\|_1 \rightarrow 0$ as $n \rightarrow \infty$.

(b) Follows immediately from (a).

◄

Exercise 9.1.13 We have seen that \mathcal{L}^1 -convergence need not imply μ -a.e. convergence, and vice versa. Show that if $(f_n)_n$ converges both in \mathcal{L}^1 and μ -a.e., then the limits are the same, i.e. show that if $f_n \rightarrow f$ in \mathcal{L}^1 and $f_n \rightarrow g$ μ -a.e., then $f = g$ μ -a.e.

[Hint: If $f_n \rightarrow f$ in \mathcal{L}^1 , we may choose a subsequence so that $f_{n_k} \rightarrow f$ μ -a.e.]

◻

Now we look at $\mathcal{L}^2(\Omega, \mathcal{F}, \mu) := \{f : \Omega \rightarrow \mathbb{C} : f \text{ is } \mathcal{F}\text{-measurable, and } \|f\|_2 < \infty\}$, where $\|f\|_2 := (\int |f|^2 d\mu)^{\frac{1}{2}}$. This space is nicer than \mathcal{L}^1 . Not only is it a Banach space: it is a Hilbert space, and therefore we can do geometry there.

Define a map $\langle \cdot, \cdot \rangle : \mathcal{L}^2 \times \mathcal{L}^2 \rightarrow \mathbb{C}$ by

$$\langle f, g \rangle =: \int f \bar{g} d\mu$$

We will show that this is an inner product, provided we agree that two functions are the same when they are μ -a.e. equal.

There are some technical details that need to be verified before we can proceed:

Lemma 9.1.14 *If $f, g \in \mathcal{L}^2$, then $fg \in \mathcal{L}^1$.*

Proof: As $(|f| - |g|)^2 \geq 0$, we have that $|fg| \leq 2|f| |g| \leq |f|^2 + |g|^2$, and thus that $\int |fg| d\mu \leq \int |f|^2 d\mu + \int |g|^2 d\mu < \infty$ when $f, g \in \mathcal{L}^2$.

◄

Since $|g|^2 = |\bar{g}|^2$, we see that $f, g \in \mathcal{L}^2$ implies $f, \bar{g} \in \mathcal{L}^2$, which in turn implies that $f\bar{g} \in \mathcal{L}^1$. Hence $\langle f, g \rangle := \int f \bar{g} d\mu$ exists when $f, g \in \mathcal{L}^2$.

Lemma 9.1.15 *The map $\langle \cdot, \cdot \rangle : \mathcal{L}^2 \times \mathcal{L}^2 \rightarrow \mathbb{C} : (f, g) \mapsto \int f \bar{g} d\mu$ is an inner product on $\mathcal{L}^2(\Omega, \mathcal{F}, \mu)$, provided we agree that two functions are the same when they are μ -a.e. equal, i.e.*

- (i) $\langle f_1 + f_2, g \rangle = \langle f_1, g \rangle + \langle f_2, g \rangle$
- (ii) $\langle cf, g \rangle = c\langle f, g \rangle$ for all $c \in \mathbb{C}$.
- (iii) $\langle f, g \rangle = \overline{\langle g, f \rangle}$
- (iv) $\langle f, f \rangle \geq 0$, and $\langle f, f \rangle = 0$ if and only if $f = 0$ μ -a.e.

Exercise 9.1.16 Prove Lemma 9.1.15.

□

We already know that for real spaces, the inner product induces a norm, defined by

$$\|v\| = \langle v, v \rangle^{\frac{1}{2}}$$

We want to verify that this is also the case for complex spaces.

Note that for any $\lambda \in \mathbb{C}$ we have

$$\begin{aligned} 0 \leq \langle v - \lambda w, v - \lambda w \rangle &= \langle v, v \rangle - \langle v, \lambda w \rangle - \langle \lambda w, v \rangle + \langle \lambda w, \lambda w \rangle \\ &= \langle v, v \rangle + \bar{\lambda} \langle v, w \rangle + \lambda \overline{\langle v, w \rangle} + |\lambda|^2 \langle w, w \rangle \\ &= \langle v, v \rangle + 2\operatorname{Re}(\lambda \langle v, w \rangle) + |\lambda|^2 \langle w, w \rangle \\ &\leq \|v\|^2 + 2|\lambda| \cdot |\langle v, w \rangle| + |\lambda|^2 \|w\|^2 \end{aligned}$$

Thus

$$\|v\|^2 + 2|\lambda| \cdot |\langle v, w \rangle| + |\lambda|^2 \|w\|^2 \geq 0 \quad \text{for all } \lambda \in \mathbb{C} \quad (*)$$

This is a quadratic polynomial in $|\lambda|$ which is always non-negative. Thus, noting that the discriminant must be ≤ 0 , or else substituting $\lambda := \frac{\langle v, w \rangle}{\langle w, w \rangle}$, we obtain:

Theorem 9.1.17 (Cauchy–Schwarz Inequality)

In any inner product space, $(V, \langle \cdot, \cdot \rangle)$, we have

$$|\langle v, w \rangle| \leq \|v\| \cdot \|w\|$$

where $\|v\| := \langle v, v \rangle^{\frac{1}{2}}$.

Lemma 9.1.18 If V is a vector space over \mathbb{C} , and that $\langle \cdot, \cdot \rangle : V \times V \rightarrow \mathbb{C}$ is an inner product on V , i.e. a map satisfying (i)–(iv) Lemma 9.1.15. Define $\|\cdot\| : V \rightarrow \mathbb{R}_+$ by

$$\|v\| = \langle v, v \rangle^{\frac{1}{2}}$$

Then $\|\cdot\|$ is a norm on V , i.e.

- (i) $\|v + w\| \leq \|v\| + \|w\|$
- (ii) $\|cv\| = |c| \cdot \|v\|$ for all $c \in \mathbb{C}$
- (iii) $\|v\| \geq 0$, and $\|v\| = 0$ if and only if $v = 0$.

Proof: Using the Cauchy–Schwarz inequality, we see that have

$$\begin{aligned} \|v + w\|^2 &= \langle v + w, v + w \rangle = \|v\|^2 + 2\operatorname{Re}(\langle v, w \rangle) + \|w\|^2 \leq \|v\|^2 + 2|\langle v, w \rangle| + \|w\|^2 \\ &\leq \|v\|^2 + 2\|v\| \cdot \|w\| + \|w\|^2 \\ &= (\|v\| + \|w\|)^2 \end{aligned}$$

which yields the triangle inequality (i).

(ii), (ii) are left as easy exercises.

–

Corollary 9.1.19 *The map $\|\cdot\|_2 : \mathcal{L}^2 \mapsto \mathbb{R} : f \mapsto \langle f, f \rangle^{\frac{1}{2}} = (\int |f|^2 d\mu)^{\frac{1}{2}}$ is a norm on $\mathcal{L}^2(\Omega, \mathcal{F}, \mu)$, induced by the inner product.*

We also have:

Theorem 9.1.20 (Hölder’s inequality for \mathcal{L}^2)

If $f, g \in \mathcal{L}^2(\Omega, \mathcal{F}, \mu)$, then

$$\|fg\|_1 \leq \|f\|_2 \|g\|_2$$

Exercise 9.1.21 Use the Cauchy–Schwarz inequality to prove Thm. 9.1.20.

□

Associated with $\|\cdot\|_2$ is another notion of convergence, namely *convergence in mean square* or \mathcal{L}^2 –convergence. If $f_n, f \in \mathcal{L}^2(\Omega, \mathcal{F}, \mu)$ for $n \in \mathbb{N}$, then we say that

$$f_n \xrightarrow{\mathcal{L}^2} f \quad \text{if and only if} \quad \|f_n - f\|_2 \rightarrow 0$$

Exercise 9.1.22 (a) On $((0, 1], \mathcal{B}((0, 1]), \lambda)$, define $f_n = I_{A_n}$, where $A_{2^n+k} := (\frac{k}{2^n}, \frac{k+1}{2^n}]$ for $n, k \in \mathbb{N}$, $0 \leq k < 2^n$. Show that $f_n \rightarrow 0$ in \mathcal{L}^2 , but that (f_n) does not converge a.e.

(b) On $(\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+), \lambda)$ define $f_n := \sum_{k=1}^n \frac{1}{k} I_{(k-1, k]}$. Show that $(f_n)_n$ converges in \mathcal{L}^2 , but not in \mathcal{L}^1 .

(c) On $((0, 1], \mathcal{B}((0, 1]), \lambda)$ define $f_n := \sum_{k=1}^n \sqrt{k} I_{(\frac{1}{k+1}, \frac{1}{k}]}$. Show that $(f_n)_n$ converges in \mathcal{L}^1 but not in \mathcal{L}^2 .

□

The preceding exercise shows that \mathcal{L}^2 convergence need not imply \mathcal{L}^1 –convergence, or vice versa. For probability spaces, however, we have the following:

Theorem 9.1.23 *If $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, then*

$$\|f\|_1 \leq \|f\|_2$$

Thus $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P}) \subseteq \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$. Moreover, if $f_n \rightarrow f$ in \mathcal{L}^2 , then also $f_n \rightarrow f$ in \mathcal{L}^1 .

Exercise 9.1.24 Prove Thm. 9.1.23.

[Hint: Show $1 \in \mathcal{L}^2$. Apply Hölder’s inequality to f and 1.]

□

Recall that an inner product space which is complete (i.e. in which every Cauchy sequence converges) is called a Hilbert space.

Theorem 9.1.25 (Riesz–Fischer Theorem for \mathcal{L}^2)
 $\mathcal{L}^2(\Omega, \mathcal{F}, \mu)$ is complete w.r.t. to the norm $\|\cdot\|_2$ induced by the inner product. Thus it is a Hilbert space.

Proof: For $k \in \mathbb{N}$, choose an increasing sequence $(n_k)_k$ of natural no. such that $\sup_{m \geq n_k} \|f_m - f_{n_k}\|_2 < 2^{-k}$. Then by the MCT $\|\sum_k |f_{n_{k+1}} - f_{n_k}|\|_2 \leq \sum_k \|f_{n_{k+1}} - f_{n_k}\|_2 < \infty$. hence $\sum_k |f_{n_{k+1}} - f_{n_k}| < \infty$ μ -a.e., and hence $(f_{n_k})_k$ is a Cauchy sequence μ -a.e. Define $f : \Omega \rightarrow \mathbb{R}$ by $f(\omega) = \lim_k f_{n_k}(\omega)$, if this limit exists, and $f(\omega) = 0$ else. Then f is measurable, and that $f_{n_k} \rightarrow f$ μ -a.e. as $k \rightarrow \infty$. By Fatou's Lemma,

$$\|f\|_2 \leq \liminf_n \|f_n\|_2 < \infty$$

(because Cauchy sequences are bounded), so that $f \in \mathcal{L}^2$, and similarly

$$\|f - f_n\|_2 = \liminf_k \|f_{n_k} - f_n\|_2 \leq \sup_{m \geq n} \|f_m - f_n\|_2 \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

Thus $f_n \xrightarrow{\mathcal{L}^2} f$.

—

9.1.4 General Theory of \mathcal{L}^p -spaces*

We recall here the definition of the \mathcal{L}^p -spaces. We also introduce the space \mathcal{L}^∞ :

Definition 9.1.26 Suppose that $(\Omega, \mathcal{F}, \mu)$ is a measure space. If $1 \leq p < \infty$ (p need not be an integer), then $\mathcal{L}^p(S, \mathcal{F}, \mu)$ is defined to be the set of all \mathcal{F} -measurable (real- or complex-valued) functions $\Omega \xrightarrow{f} \mathbb{R}$ such that $|f|^p$ is μ -integrable.

A function $\Omega \xrightarrow{f} \mathbb{R}$ is said to be *essentially bounded* iff there is a real number M such that $|f| \leq M$ μ -a.e.

$\mathcal{L}^\infty(\Omega, \mathcal{F}, \mu)$ is defined to be the set of all essentially bounded \mathcal{F} -measurable $\Omega \xrightarrow{f} \mathbb{R}$.

For each $1 \leq p < \infty$, we define a map $\|\cdot\|_p : \mathcal{L}^p(\Omega, \mathcal{F}, \mu) \rightarrow \mathbb{R}$ by

$$\|f\|_p := (\mu|f|^p)^{\frac{1}{p}}$$

We also define a map $\|\cdot\|_\infty : \mathcal{L}^\infty(\Omega, \mathcal{F}, \mu) \rightarrow \mathbb{R}$ by

$$\|f\|_\infty = \inf\{M : |f| \leq M \text{ } \mu\text{-a.e.}\}$$

The maps $\|\cdot\|_p$ are called \mathcal{L}^p -norms, or just p -norms.

Remarks 9.1.27 The definition of $\mathcal{L}^\infty(\Omega, \mathcal{F}, \mu)$ differs from that of the other \mathcal{L}^p -spaces, so it is worth elaborating a little on it. A real- or complex-valued measurable function f is *essentially bounded* if there is a real number $M \geq 0$ such that $|f| \leq M$ μ -a.e., i.e. the set $\{\omega \in \Omega : |f(\omega)| > M\}$ is a μ -null set. Call such

an M an *essential bound* of f . Then $\|f\|_\infty$ is defined to be the infimum of all the essential bounds. Note that $\|f\|_\infty$ is itself an essential bound of f . Indeed,

$$\{\omega \in \Omega : |f(\omega)| > \|f\|_\infty\} = \bigcup_n \{\omega \in \Omega : |f(\omega)| > \|f\|_\infty + \frac{1}{n}\}$$

is a countable union of μ -null sets, and thus itself a μ -null set. Hence $\|f\|_\infty$ is the smallest essential bound of f , i.e. for all $M < \|f\|_\infty$, we have $\mu\{\omega \in \Omega : |f(\omega)| > M\} > 0$.

We have already shown that the \mathcal{L}^p -spaces are vector spaces for $1 \leq p < \infty$. The same is true for \mathcal{L}^∞ : If $f, g \in \mathcal{L}^\infty$, then $|f(\omega) + g(\omega)| \leq |f(\omega)| + |g(\omega)| \leq \|f\|_\infty + \|g\|_\infty$ for μ -almost all $\omega \in \Omega$. hence $\|f\|_\infty + \|g\|_\infty$ is an essential bound for $f + g$, and moreover, we have a triangle inequality:

$$\|f + g\|_\infty \leq \|f\|_\infty + \|g\|_\infty$$

□

If the underlying measure space is understood from context, we shall write \mathcal{L}^p instead of $\mathcal{L}^p(S, \mathcal{S}, \mu)$. For the next theorem, note that if $a, b \geq 0$, and if $1 < p, q < \infty$ are such that $p^{-1} + q^{-1} = 1$, then

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}$$

To see this, define $h(t) = tb - \frac{t^p}{p}$, and find the maximum of h . Alternatively, apply the Arithmetic–Geometric Mean inequality.

Theorem 9.1.28 *Let (S, \mathcal{S}, μ) be a measure space and let f, g be real-valued \mathcal{S} -measurable functions.*

(a) **HÖLDER'S INEQUALITY:** *Suppose that $1 \leq p \leq \infty$ and that $p^{-1} + q^{-1} = 1$. If $f \in \mathcal{L}^p$, and $g \in \mathcal{L}^q$, then $fg \in \mathcal{L}^1$, and*

$$\|fg\|_1 \leq \|f\|_p \|g\|_q$$

(b) **MINKOWSKI'S INEQUALITY:** *Let $p \geq 1$. If $f, g \in \mathcal{L}^p$, then*

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p$$

Proof: (a) If $p = 1$ (so that $q = +\infty$), then $|fg| \leq |f| \|g\|_\infty$ μ -a.e. and so $\|fg\|_1 = \mu|fg| \leq \mu|f| \cdot \|g\|_\infty = \|f\|_1 \|g\|_\infty < \infty$.

If $p > 1$, put $a = \frac{|f(s)|}{\|f\|_p}$, $b = \frac{|g(s)|}{\|g\|_q}$ and apply the remark just before the statement of the theorem to conclude

$$\frac{|f(s)g(s)|}{\|f\|_p \|g\|_q} \leq \frac{|f(s)|^p}{p \|f\|_p^p} + \frac{|g(s)|^q}{q \|g\|_q^q}$$

Integrating both sides w.r.t. μ yields the result.

(b) This relation is easy to prove if $p = 1$ or $p = \infty$. For $1 < p < \infty$, note that $q = \frac{p}{p-1}$, and thus that $|f + g|^{p-1} \in \mathcal{L}^q$. By Hölder's inequality,

$$\begin{aligned} \|f + g\|_p^p &\leq \int |f| |f + g|^{p-1} d\mu + \int |g| |f + g|^{p-1} d\mu \\ &\leq \|f\|_p \cdot \|(f + g)^{p-1}\|_q + \|g\|_p \cdot \|(f + g)^{p-1}\|_q \\ &= (\|f\|_p + \|g\|_p) \|f + g\|_p^{p-1} \end{aligned}$$

⊣

It is now clear that $\|\cdot\|_p$ satisfies the following:

- (i) $\|f\|_p \geq 0$ $\|f\|_p = 0$ iff $f = 0$ μ -a.e.
- (ii) If $\alpha \in \mathbb{R}$, then $\|\alpha f\|_p = |\alpha| \|f\|_p$.
- (iii) $\|f + g\|_p \leq \|f\|_p + \|g\|_p$

Thus $\|\cdot\|_p$ is *almost* a norm on \mathcal{L}^p . The requirement that $\|f\|_p = 0$ iff $f = 0$ doesn't hold, but holds only almost everywhere. To get a *bona fide* norm, we must identify any two functions that are equal μ -a.e:

Definition and Proposition 9.1.29 Let (S, \mathcal{S}, μ) be a measure space, and let $1 \leq p \leq \infty$. Define a relation \equiv on \mathcal{L}^p by $f \equiv g$ iff $f = g$ μ -a.e. Then \approx is an equivalence relation on \mathcal{L}^p . Define $[f] := \{g \in \mathcal{L}^p : f \equiv g\}$. Then $[0] = \{g \in \mathcal{L}^p : g = 0 \text{ } \mu\text{-a.e.}\}$ is a vector subspace of \mathcal{L}^p , and $[f] = f + [0] := \{f + g : g \in [0]\}$. Let

$$L^p(S, \mathcal{S}, \mu) = \{[f] : f \in \mathcal{L}^p(S, \mathcal{S}, \mu)\}$$

Then L^p is a vector space and the map, which by abuse of notation we also call $\|\cdot\|_p$, which is defined by

$$\|[f]\|_p := \|f\|_p$$

is a norm on L^p .

Proof: That \equiv is an equivalence relation is straightforward, as is the statement that $[0]$ is a vector subspace of \mathcal{L}^p . It is also easy to see that L^p is a vector space, if the operations are defined in the natural way (e.g. $[f] + [g] := [f + g]$ — one must check that this is well-defined, i.e. that if $[f_1] = [f_2]$ and $[g_1] = [g_2]$, then $[f_1 + g_1] = [f_2 + g_2]$, but that is easy.) $[0]$ is clearly the zero vector in L^p . Also, if $[f_1] = [f_2]$, then $f_1 = f_2$ μ -a.e., and thus $\mu f_1^p = \mu f_2^p$, which shows that $\|f_1\|_p = \|f_2\|_p$ (in case $p < \infty$), and thus that $\|\cdot\|_p$ is well-defined on L^p . To see that it is a norm, note that (i) $\|[f]\|_p = \|f\|_p \geq 0$, and that $\|[f]\|_p = 0$ iff $f = 0$ μ -a.e. iff $[f] = [0]$; (ii) $\|\alpha[f]\|_p = \|\alpha f\|_p = |\alpha| \|f\|_p$, and (iii) $\|[f] + [g]\|_p = \|f + g\|_p \leq \|f\|_p + \|g\|_p$.

In case $p = \infty$, it is also straightforward to see that $\|\cdot\|_\infty$ is a well-defined norm on L^∞ .

⊣

In practice, we usually don't bother too much about the distinction between \mathcal{L}^p and L^p .

Now that we know that $\|\cdot\|_p$ is a norm, we have a notion of convergence:

Definition 9.1.30 A sequence $(f_n)_n$ in $\mathcal{L}^p(S, \mathcal{S}, \mu)$ is said to converge to $f \in \mathcal{L}^p(S, \mathcal{S}, \mu)$ in L^p (or in p^{th} mean) iff $\|f_n - f\|_p \rightarrow 0$.

□

We now have two notions of convergence for measurable functions: almost everywhere convergence, and convergence in mean. We write

$$f_n \xrightarrow{\text{a.e.}} f \quad f_n \xrightarrow{L^p} f$$

In a later section, we will investigate this, and other, notions of convergence in greater detail.

Theorem 9.1.31 (Riesz–Fischer)

If (S, \mathcal{S}, μ) is a measure space and $1 \leq p \leq \infty$, then $L^p(S, \mathcal{S}, \mu)$ is a Banach space.

Proof: Suppose that $(f_n)_n$ is a Cauchy sequence in \mathcal{L}^p , i.e. that $\sup_{m \geq n} \|f_m - f_n\|_p \rightarrow 0$ as $n \rightarrow \infty$.

First assume that $p > 1$. For $k \in \mathbb{N}$, choose an increasing sequence $(n_k)_k$ of natural no. such that $\sup_{m \geq n_k} \|f_m - f_{n_k}\|_p < 2^{-k}$. Then by the MCT $\|\sum_k |f_{n_{k+1}} - f_{n_k}|\|_p \leq \sum_k \|f_{n_{k+1}} - f_{n_k}\|_p < \infty$. hence $\sum_k |f_{n_{k+1}} - f_{n_k}| < \infty$ μ -a.e., and hence $(f_{n_k})_k$ is a Cauchy sequence μ -a.e. Define $f : S \rightarrow \mathbb{R}$ by $f(s) = \lim_k f_{n_k}(s)$, if this limit exists, and $f(s) = 0$ else. Then f is measurable, and that $f_{n_k} \rightarrow f$ μ -a.e. as $k \rightarrow \infty$. Then by Fatou's Lemma,

$$\|f\|_p \leq \liminf_n \|f_n\|_p < \infty$$

(because Cauchy sequences are bounded), so that $f \in \mathcal{L}^p$, and similarly

$$\|f - f_n\|_p = \liminf_k \|f_{n_k} - f_n\|_p \leq \sup_{m \geq n} \|f_m - f_n\|_p \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

Thus $f_n \xrightarrow{L^p} f$.

Next, assume that $p = \infty$. We have $\sup_{m \geq n} |f_m - f_n| \leq \sup_{m \geq n} \|f_m - f_n\|_\infty$ μ -a.e., and thus $(f_n)_n$ is a Cauchy sequence μ -a.e. Define $f : S \rightarrow \mathbb{R}$ as above: $f(s) = \lim_n f_n(s)$ if this limit exists, and $f(s) = 0$ otherwise. Then

$$|f| \leq |f_n| + |f_n - f| = |f_n| + \lim_m |f_n - f_m| \leq \|f_n\|_\infty + \sup_{m \geq n} \|f_n - f_m\|_\infty \quad \mu\text{-a.e.}$$

so that $f \in \mathcal{L}^\infty$, and

$$|f_n - f| = \lim_k |f_n - f_k| \leq \sup_{m \geq n} \|f_n - f_m\|_\infty \quad \mu\text{-a.e.}$$

Hence $\|f_n - f\|_\infty \leq \sup_{m \geq n} \|f_n - f_m\|_\infty \rightarrow 0$ as $n \rightarrow \infty$, proving that $f_n \xrightarrow{L^\infty} f$.

□

The following result is easy:

Theorem 9.1.32 Let (S, \mathcal{S}, μ) be a measure space. The map

$$\langle \cdot, \cdot \rangle : \mathcal{L}^2 \times \mathcal{L}^2 \rightarrow \mathbb{R} : (f, g) \mapsto \int fg \, d\mu$$

is an inner product on \mathcal{L}^2 which induces the L^2 -norm $\|\cdot\|_2$. Hence L^2 is a Hilbert space.

□

Exercise 9.1.33 Prove Thm 9.1.32.

□

For probability theory, the following result is also useful:

Proposition 9.1.34 Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. If $1 \leq p \leq r \leq \infty$, then

$$\|X\|_p \leq \|X\|_r$$

for any random variable X , so that $\mathcal{L}^r(\Omega, \mathcal{F}, \mathbb{P}) \subseteq \mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$.

Moreover, if $X \in \mathcal{L}^\infty$, then

$$\|X\|_\infty = \lim_{p \rightarrow \infty} \|X\|_p$$

Proof: Note that if $X \in \mathcal{L}^r$, then $X^p \in \mathcal{L}^{\frac{r}{p}}$. Now $p' = \frac{r}{p}$ and $q' = \frac{r}{r-p}$, satisfy the relation $\frac{1}{p'} + \frac{1}{q'} = 1$, and so Hölder's inequality applied to $f = |X|^p$ and $g = 1$ yields

$$\|X\|_p^p = \int f g \, d\mathbb{P} \leq \|f\|_{p'} \cdot \|g\|_{q'} = \left(\int |X^p|^{\frac{r}{p}} \, d\mathbb{P} \right)^{\frac{p}{r}} \cdot 1 = \|X\|_r^p$$

Next, suppose that $X \in \mathcal{L}^\infty$. Then $\|X\|_p \leq \|X\|_\infty$, so $\limsup_p \|X\|_p \leq \|X\|_\infty$.

If $M < \|X\|_\infty$, then $\int |X|^p \, d\mathbb{P} \geq M^p \mathbb{P}(|X| > M)$ and so $\|X\|_p \geq M \mathbb{P}(|X| > M)^{\frac{1}{p}}$. Now $\mathbb{P}(|X| > M) > 0$, because $M < \|X\|_\infty$, and thus $\liminf_p \|X\|_p \geq M$ (because $\mathbb{P}(|X| > M)^{\frac{1}{p}} \rightarrow 1$). Since M was arbitrary, also $\liminf_p \|X\|_p \geq \|X\|_\infty$.

◄

9.2 Geometry of Hilbert Space and Generalized Fourier Series

9.2.1 Projections in Hilbert Spaces

We have already studied some Hilbert space theory earlier in this course, and we will repeat here the most important facts that were then obtained.

Recall that in \mathbb{R}^n , the dot product does not only induce a length (i.e. a norm), but also an *angle*: The angle θ between two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ is given by

$$\cos \theta = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

We can imitate this definition in an abstract inner product space $(V, \langle \cdot, \cdot \rangle)$, and define the angle between $x, y \in V$ by

$$\cos \theta := \frac{\langle x, y \rangle}{\|x\| \|y\|} \quad \text{where } \|x\| := \sqrt{\langle x, x \rangle}$$

By the Cauchy–Schwarz inequality it follows immediately that $|\cos \theta| \leq 1$, so that this definition makes sense. It also follows that $|\cos \theta| = 1$ if and only if x is a scalar multiple of y , i.e. iff x, y are parallel. We don't really need the concept of angle, but we do want the associated concept of *orthogonality*:

Definition 9.2.1 Suppose that $(V, \langle \cdot, \cdot \rangle)$ is an inner product space. We say that $x, y \in V$ are *orthogonal*, and write $x \perp y$, if and only if $\langle x, y \rangle = 0$. If $G \subseteq V$, we say that $x \perp G$ iff $\forall g \in G (x \perp g)$.

First, you may prove the easy

Proposition 9.2.2 *If V is a Hilbert space and $v, w \in V$, then*

(a) (Parallellogram Law) $\|v - w\|^2 + \|v + w\|^2 = 2(\|v\|^2 + \|w\|^2)$

(b) (Pythagoras) *If $v \perp w$, then $\|v + w\|^2 = \|v\|^2 + \|w\|^2$.*

If V is a linear subspace of \mathbb{R}^n , then we can project any $\mathbf{x} \in \mathbb{R}^n$ onto V . That is, we can represent \mathbf{x} as a sum

$$\mathbf{x} = \mathbf{x}^{\parallel} + \mathbf{x}^{\perp} \quad \text{where } \mathbf{x}^{\parallel} \in V, \quad \mathbf{x}^{\perp} \perp V$$

One can think of \mathbf{x}^{\parallel} as the best approximation to \mathbf{x} in V : It is the vector in V which lies closest to \mathbf{x} .

Suppose that V is a Hilbert space, and that W is a closed linear subspace of V . If $v_0 \in V$, we can find the *best approximation* of v_0 in W . This is the unique vector w_0 with the properties that

(i) $w_0 \in W$, and

(ii) $\|v_0 - w_0\| = \inf\{\|v_0 - w\| : w \in W\}$, i.e. w_0 is the vector in W that lies closest to v_0 .

(iii) Moreover, $(v_0 - w_0) \perp W$.

The vector w_0 satisfying (i)–(iii) is called the *orthogonal projection of v_0 onto W* . Indeed, $v_0 = w_0 + (v_0 - w_0)$ decomposes v_0 into a vector in W and a vector orthogonal to W . It remains to show that orthogonal projections exist and are unique.

Proposition 9.2.3 *Let V be a Hilbert space, and let W be a closed linear subspace of V . Then any v_0 in V has a unique decomposition*

$$v_0 = v_0^{\parallel} + v_0^{\perp} \quad \text{where } v_0^{\parallel} \in W, \quad v_0^{\perp} \perp W$$

v_0^{\parallel} is called the orthogonal projection of v_0 onto W .

Proof: *Uniqueness:* If

$$v_0 = v_0^{\parallel} + v_0^{\perp} = u_0^{\parallel} + u_0^{\perp}$$

where $v_0^{\parallel}, u_0^{\parallel} \in W$ and $v_0^{\perp}, u_0^{\perp} \perp W$, then

$$v_0^{\parallel} - u_0^{\parallel} = u_0^{\perp} - v_0^{\perp} =: x$$

is a vector with the properties that $x \in W$ and that $x \perp W$. This implies that $x \perp x$, i.e. that $\langle x, x \rangle = 0$. Hence $x = 0$, and so $v_0^{\parallel} = u_0^{\parallel}$, $v_0^{\perp} = u_0^{\perp}$.

Existence: Let $\delta = \inf\{\|v_0 - w\| : w \in W\}$, and choose a sequence $w_n \in W$ such that $\|v_0 - w_n\| \rightarrow \delta$. We show that $(w_n)_n$ is a Cauchy sequence in W : for if $\varepsilon > 0$, we may choose N such that $\|v_0 - w_n\|^2 - \delta^2 < \varepsilon$ whenever $n \geq N$. By the Parallelogram Law it follows that if $n, m \geq N$, then

$$2\varepsilon + 2\delta^2 > \|v_0 - w_n\|^2 + \|v_0 - w_m\|^2 = 2\|v_0 - \frac{1}{2}(w_n + w_m)\|^2 + 2\|\frac{1}{2}(w_n - w_m)\|^2 \geq 2\delta^2 + \frac{1}{2}\|w_n - w_m\|^2$$

Since $(w_n)_n$ is a Cauchy sequence, and since W is closed, there is $w_0 \in W$ such that $w_n \rightarrow w_0$. We will show that $w_0 = v_0^{\parallel}$. The fact that $\|v_0 - w_0\| \leq \|v_0 - w_n\| + \|w_n - w_0\|$ (for all $n \in \mathbb{N}$) then is easily seen to imply that $\|v_0 - w_0\| = \delta$.

It remains to show that $v_0 - w_0 \perp W$. Given an arbitrary $w \in W$ and a real $\lambda \in \mathbb{R}$, have $\|v_0 - w_0\|^2 = \delta^2 \leq \|v_0 - (w_0 + \lambda w)\|^2$, so that

$$-2\lambda \langle v_0 - w_0, w \rangle + \lambda^2 \|w\|^2 \geq 0$$

Since this holds for all real λ we must have $\langle v_0 - w_0, w \rangle = 0$. (Another way to see this is to note that the quadratic in λ has a unique root at $\lambda = 0$) and to calculate the discriminant.)

—

Exercise 9.2.4 Consider again the problem of linear least squares estimation: Given functions $\phi_1(t), \dots, \phi_n(t)$ and observations $(t_1, y_1), \dots, (t_m, y_m)$, we seek coefficients x_1, \dots, x_n which *best* fit the data, i.e. so that

$$y_j \approx \sum_{i=1}^n x_i \phi_i(t_j) \quad \text{for } j = 1, \dots, m$$

Here, “best” means the following: For each $\mathbf{x} := (x_1, \dots, x_n)$ we obtain numbers $y_j^*(\mathbf{x}) := \sum_{i=1}^n x_i \phi_i(t_j)$. We want that $y_j \approx y_j^*(\mathbf{x})$ for all $j = 1, \dots, m$, i.e. we want to determine that n -tuple \mathbf{x} for which the combined error is as small as possible. We therefore want the *distance* between the vectors $\mathbf{y} := (y_1, \dots, y_m)$ and $\mathbf{y}^*(\mathbf{x}) := (y_1^*, \dots, y_m^*)$ to be as small as possible, i.e. we want to find that optimal value \mathbf{x}^* of \mathbf{x} for which

$$\|\mathbf{y} - \mathbf{y}^*(\mathbf{x})\|^2 = \sum_{j=1}^m (y_j - y_j^*(\mathbf{x}))^2$$

is a minimum, where $\|\cdot\|$ denotes the usual Euclidean norm on \mathbb{R}^m .

- (a) On $(\mathbb{R}, \mathcal{P}(\mathbb{R}))$, consider the measure $\mu = \delta_{t_1} + \dots + \delta_{t_m}$ (where δ_a is the Dirac measure located at a). For $f : \mathbb{R} \rightarrow \mathbb{R}$, show that $\|f\|_2^2 = \sum_{j=1}^m |f(t_j)|^2$.
- (b) Conclude that $\mathcal{L}^2(\mathbb{R}, \mathcal{P}(\mathbb{R}), \mu)$ is simply the set of *all* functions $f : \mathbb{R} \rightarrow \mathbb{R}$.
- (c) We need, however, to recall the convention that we regard functions which are μ -a.e. equal as the same function. Show that if $f, g : \mathbb{R} \rightarrow \mathbb{R}$, then

$$f = g \quad \mu\text{-a.e.} \quad \text{iff} \quad f(t_j) = g(t_j) \quad \text{for all } j = 1, \dots, m$$

- (d) Suppose now that we are given functions $\phi_1(t), \dots, \phi_n(t)$. Show that

$$W := \left\{ \sum_{i=1}^n x_i \phi_i : x_1, \dots, x_n \in \mathbb{R} \right\}$$

is a closed linear subspace of \mathcal{L}^2 .

- (e) Given $(t_1, y_1), \dots, (t_m, y_m)$. Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be any function such that $f(t_j) = y_j$ for all $j = 1, \dots, m$. (Any two such functions are μ -a.e. equal, and thus “the same”). Let ψ be the orthogonal projection of f onto W . Explain why there exist $x_1^*, \dots, x_n^* \in \mathbb{R}$ such that $\psi = \sum_{i=1}^n x_i^* \phi_i$. Now explain why $\mathbf{x}^* = (x_1^*, \dots, x_n^*)$ is precisely the solution to the least squares estimation problem.

Thus:

Least Squares Estimation = Orthogonal Projection

□

Exercise 9.2.5 Suppose that W is a closed subspace of a Hilbert space $(V, \langle \cdot, \cdot \rangle)$. For each $v \in V$, let v^{\parallel} be the orthogonal projection of v onto W . Define a map $P : V \rightarrow V$ by $Pv := v^{\parallel}$. Use the uniqueness of the orthogonal projection to prove the following results.

- (a) Show that P is a bounded linear operator.
- (b) Show that P is *idempotent*, i.e. that $P^2 = P$ (i.e. $P(Pv) = Pv$ for all $v \in V$).
- (c) Show that P is *self-adjoint*, i.e. that $\langle Pv_1, v_2 \rangle = \langle Pv_1, Pv_2 \rangle = \langle v_1, Pv_2 \rangle$ for all $v_1, v_2 \in V$.
- (d) Show that $\ker P = (\operatorname{ran} P)^{\perp} = W^{\perp}$.

□

9.2.2 Orthonormal Bases

It is well-known that every vector space has a basis, i.e. a maximal linearly independent set. We call such a basis a *Hamel basis*. Every vector can then be written as a linear combination of these basis vectors in a unique way.

For Hilbert spaces, we have another notion of basis, namely that of *orthonormal basis*.

Definition 9.2.6 A subset Φ of a Hilbert space V is set to be an *orthonormal subset* if and only

- (i) $\|\phi\| = 1$ for all $\phi \in \Phi$ (normality)
- (ii) If $\phi_1 \neq \phi_2 \in \Phi$, then $\phi_1 \perp \phi_2$ (orthogonality)

An *orthonormal basis* for V is a *maximal* orthonormal subset of V . Such a set is also referred to as a *complete orthonormal set*.

In \mathbb{R}^n , the standard basis vectors form an orthonormal basis.

Clearly, Φ is an orthonormal basis for V if and only if we cannot find any vectors v of unit length which are orthogonal to every vector ϕ in Φ . Since non-zero vectors can always be scaled to vectors of unit length, we have shown the following trivial but useful fact:

Lemma 9.2.7 Φ is an orthonormal basis for a Hilbert space V if and only if for any $v \in V$ we have

$$v \perp \phi \text{ for all } \phi \in \Phi \quad \text{implies} \quad v = 0$$

It can be shown that if Ψ is an orthonormal subset of V , then there is an orthonormal basis Φ such that $\Psi \subseteq \Phi$, i.e. that:

Proposition 9.2.8 Every orthonormal subset of a Hilbert space can be extended to an orthonormal basis.

We will not prove this proposition, but the idea is simple: Keep adding orthonormal vectors to Ψ until you can't find any more. Then stop. We did precisely this by induction earlier in the course for finite-dimensional Hilbert spaces — Gram-Schmidt orthogonalization. The only problem for general, non-finite dimensional Hilbert spaces is that one will not stop in finite time, i.e. one needs to use a *transfinite induction*, and this requires a deeper understanding of set theory than you currently possess. For this reason, we will restrict ourselves to *separable* Hilbert spaces.

Definition 9.2.9 A Hilbert space is said to be *separable* if and only if it has a countable orthonormal basis $\Phi := \{\phi_n : n \in \mathbb{N}\}$. Thus the sequence $(\phi_n)_n$ has the following properties:

- (i) $\langle \phi_n, \phi_m \rangle = \delta_{nm}$ (orthonormality)
- (ii) If $v \perp \phi_n$ for all $n \in \mathbb{N}$, then $v = 0$. (completeness)

We will now show that every vector v in a separable Hilbert space V can be expressed as “infinite linear combination” of orthonormal basis vectors $(\phi_n)_n$, i.e. that

$$v = \sum_{n=1}^{\infty} c_n \phi_n \quad \text{for unique } c_n \in \mathbb{C}$$

First, however, we need to say what we mean by the expression $\sum_{n=1}^{\infty} c_n \phi_n$. This is not an infinite series with number-terms, but *vector*-terms. The definition works in any normed linear space:

Definition 9.2.10 Suppose that $(V, \|\cdot\|)$ is a normed vector space, and $v_n, v \in V$ for $n \in \mathbb{N}$. Consider the series $\sum_{n=1}^{\infty} v_n$, and define $S_m := \sum_{n=1}^m v_n$ to be the m^{th} partial sum (which is just a sum of finitely many vectors and thus a well-defined). We say that

$$\sum_{n=1}^{\infty} v_n = v \quad \text{iff} \quad \|S_m - v\| \rightarrow 0 \quad \text{as } m \rightarrow \infty$$

Speaking of limits, note that the inner product is continuous, i.e.

Lemma 9.2.11 If V is an inner product space, $v_n, v, w \in V$, and $v_n \rightarrow v$, then $\langle v_n, w \rangle \rightarrow \langle v, w \rangle$ i.e.

$$\langle \lim_n v_n, w \rangle = \lim_n \langle v_n, w \rangle$$

Exercise 9.2.12 Prove Lemma 9.2.11.

[Hint: Apply the Cauchy–Schwarz inequality to $\langle v_n - v, w \rangle$.]

□

Remarks 9.2.13 Note that the limit in $\langle \lim_n v_n, w \rangle$ is a limit of vectors in the inner product space, whereas the limit in $\lim_n \langle v_n, w \rangle$ is a limit of complex numbers.

□

Now that we know what $v = \sum_{n=1}^{\infty} c_n \phi_n$ means, we can try to determine what the *Fourier coefficients* c_n should be for a given $v \in V$. For motivation, consider \mathbb{R}^d with the Euclidean inner product and the standard basis $(\mathbf{e}_n)_{n=1,\dots,d}$. Any vector $\mathbf{x} = (x_1, \dots, x_d)$ can be written as $\mathbf{x} = \sum_{n=1}^d x_n \mathbf{e}_n$. A little calculation shows that $x_n = \langle \mathbf{x}, \mathbf{e}_n \rangle$. We try to imitate this argument in arbitrary Hilbert spaces.

Note that if it were the case that $v = \sum_{n=1}^{\infty} c_n \phi_n$, then it seems natural to argue as follows:

$$\begin{aligned} \langle v, \phi_i \rangle &= \left\langle \sum_{n=1}^{\infty} c_n \phi_n, \phi_i \right\rangle \\ &= \sum_{n=1}^{\infty} c_n \langle \phi_n, \phi_i \rangle \\ &= \sum_{n=1}^{\infty} c_n \delta_{ni} \\ &= c_i \end{aligned}$$

This argument is purely suggestive — the above operations are valid for finite sums, and we must verify that they are valid for infinite series. We begin this now.

In the following Proposition, note that $(\phi_n)_n$ need not be an orthonormal basis, just an orthonormal set.

Proposition 9.2.14 (Bessel's Inequality)

Suppose that $(\phi_n)_n$ is an orthonormal sequence in an inner product space V , and that $v \in V$. Define $c_n := \langle v, \phi_n \rangle$. Then

$$\sum_{n=1}^{\infty} |c_n|^2 = \sum_{n=1}^{\infty} |\langle v, \phi_n \rangle|^2 \leq \|v\|^2$$

Proof: Define $S_n := \sum_{k=1}^n c_k \phi_k$, and let $v_n := v - S_n$. If $1 \leq i \leq n$, then

$$\langle v_n, \phi_i \rangle = \langle v, \phi_i \rangle - \sum_{k=1}^n \langle v, \phi_k \rangle \langle \phi_k, \phi_i \rangle = \langle v, \phi_i \rangle - \langle v, \phi_i \rangle = 0$$

and hence $v_n \perp \phi_i$ for all $i = 1, \dots, n$, so that in particular $v_n \perp S_n$. By Pythagoras,

$$\begin{aligned} \|v\|^2 &= \|v_n + S_n\|^2 = \|v_n\|^2 + \|S_n\|^2 \\ &= \|v_n\|^2 + \left\langle \sum_{i=1}^n c_i \phi_i, \sum_{j=1}^n c_j \phi_j \right\rangle \\ &= \|v_n\|^2 + \sum_{i=1}^n \sum_{j=1}^n c_i \bar{c}_j \delta_{ij} \\ &= \|v_n\|^2 + \sum_{i=1}^n |c_i|^2 \geq \sum_{i=1}^n |c_i|^2 \end{aligned}$$

Now let $n \rightarrow \infty$ to obtain the result.

—

Corollary 9.2.15 *Suppose that $(\phi_n)_n$ is an orthonormal sequence in an inner product space V , and that $v \in V$ and that $c_n := \langle v, \phi_n \rangle$. Then $c_n \rightarrow 0$ as $n \rightarrow \infty$.*

Exercise 9.2.16 Prove Corollary 9.2.15.

□

Next, we show that if $(\phi_n)_n$ is an orthonormal sequence (not necessarily a basis) in a Hilbert space V , and if $v \in V$, then $\sum_{n=1}^{\infty} \langle v, \phi_n \rangle \phi_n$ always converges (though not necessarily to v), and that inner products commute with infinite sums also.

Proposition 9.2.17 (a) *If $(\phi_n)_n$ is an orthonormal sequence in a Hilbert space V , and if $v \in V$, then $\sum_{n=1}^{\infty} \langle v, \phi_n \rangle \phi_n$ converges.*

(b) *If $v = \sum_{n=1}^{\infty} c_n \phi_n$, then $\langle v, \phi_n \rangle = c_n$.*

Proof: (a) Let $c_n := \langle v, \phi_n \rangle$, $S_n := \sum_{k=1}^n c_k \phi_k$. We must show that $(S_n)_n$ converges, and for that, it suffices to show that it is a Cauchy sequence in V . Now if $n \leq m$, then

$$\|S_n - S_m\|^2 = \left\| \sum_{k=n+1}^m c_k \phi_k \right\|^2 = \sum_{k=n+1}^m |c_k|^2$$

Now $\sum_{n=1}^{\infty} |c_n|^2$ is an increasing sequence of positive real numbers. By Bessel's inequality, $\sum_{n=1}^{\infty} |c_n|^2 \leq \|v\|^2$, and hence $\sum_{n=1}^{\infty} |c_n|^2$ converges. From this we see that $(\sum_{k=1}^n |c_k|^2)_n$ is a Cauchy sequence (as a sequence of reals converges iff it is Cauchy). It follows that, for any $\varepsilon > 0$ we have $\left| \sum_{k=1}^m |c_k|^2 - \sum_{k=1}^n |c_k|^2 \right| < \varepsilon$ whenever $m \geq n$ are sufficiently large, and thus that $\|S_m - S_n\|^2 < \varepsilon$ when $m \geq n$ are sufficiently large.

(b) We have $v = \lim_n S_n$, and hence $\langle v, \phi_k \rangle = \lim_n \langle S_n, \phi_k \rangle$. But clearly

$$\langle S_n, \phi_k \rangle = \sum_{i=1}^n c_i \delta_{ik} = \begin{cases} 0 & \text{if } n < k \\ c_k & \text{if } n \geq k \end{cases}$$

so that $\lim_n \langle S_n, \phi_k \rangle = c_k$.

⊢

Theorem 9.2.18 *Suppose that $(\phi_n)_n$ is an orthonormal sequence in a Hilbert space V . The following are equivalent:*

(a) *$(\phi_n)_n$ is complete, i.e. an orthonormal basis.*

(b) *If $v \perp \phi_n$ for all n , then $v = 0$.*

(c) *If $v \in V$, then $v = \sum_{n=1}^{\infty} \langle v, \phi_n \rangle \phi_n$*

(d) *If $v, w \in V$, then $\langle v, w \rangle = \sum_{n=1}^{\infty} \langle v, \phi_n \rangle \langle \phi_n, w \rangle$*

(e) (Parseval's Identity) *If $v \in V$, then $\|v\|^2 = \sum_{n=1}^{\infty} |c_n|^2 = \sum_{n=1}^{\infty} |\langle v, \phi_n \rangle|^2$.*

Proof: (a) \Rightarrow (b) is Lemma 9.2.7.

(b) \Rightarrow (c): Let $c_n := \langle v, \phi_n \rangle$, and $w := \sum_{n=1}^{\infty} c_n \phi_n$ (which we know exists by Propn. 9.2.17). Then by Propn. 9.2.17(b)

$$\langle v - w, \phi_n \rangle = \langle v, \phi_n \rangle - c_n = 0 \quad \text{all } n \in \mathbb{N}$$

and hence $v - w = 0$.

(c) \Rightarrow (d): If $v = \sum_{n=1}^{\infty} c_n \phi_n$ and $w = \sum_{n=1}^{\infty} d_n \phi_n$, then $\langle v, \phi_n \rangle = c_n$, $\langle \phi_n, w \rangle = \overline{\langle w, \phi_n \rangle} = \bar{d}_n$ and

$$\langle v, w \rangle = \lim_{p \rightarrow \infty} \sum_{n=1}^p \langle c_n \phi_n, w \rangle = \lim_{p \rightarrow \infty} \sum_{n=1}^p \left(\lim_{q \rightarrow \infty} \sum_{m=1}^q \langle c_n \phi_n, d_m \phi_m \rangle \right) = \lim_{p \rightarrow \infty} \sum_{n=1}^p c_n \bar{d}_n = \sum_{n=1}^{\infty} c_n \bar{d}_n$$

(d) \Rightarrow (e): $\|v\|^2 = \langle v, v \rangle = \sum_{n=1}^{\infty} \langle v, \phi_n \rangle \langle \phi_n, v \rangle = \sum_{n=1}^{\infty} c_n \bar{c}_n$.

(e) \Rightarrow (a): If $(\phi_n)_n$ is not complete, i.e. not a maximal orthonormal set, then there is $\psi \in V$ so that $\|\psi\| = 1$ and such that $\psi \perp \phi_n$ for all n . Then by hypothesis

$$\|\psi\|^2 = \sum_{n=1}^{\infty} |\langle \psi, \phi_n \rangle|^2 = 0$$

contradiction.

—

We have now shown that if $(\phi_n)_n$ is a complete orthonormal sequence in a Hilbert space V , then every vector v can be written as an infinite series

$$v = \sum_{n=1}^{\infty} c_n \phi_n \quad \text{where } c_n = \langle v, \phi_n \rangle$$

The c_n are known as the Fourier coefficients of the vector v relative to the orthonormal basis $(\phi_n)_n$.

Even if the orthonormal sequence $(\phi_n)_n$ is *not* complete, the sum $\sum_{n=1}^{\infty} c_n \phi_n$ nevertheless has a nice interpretation:

Theorem 9.2.19 *Suppose that V is a Hilbert space, and that W is a closed subspace of V . Let $(\phi_n)_n$ be an orthonormal basis for W . Then*

$$\sum_{n=1}^{\infty} \langle v, \phi_n \rangle \phi_n \quad \text{is the orthogonal projection of } v \text{ onto } W$$

Exercise 9.2.20 Prove Thm. 9.2.19.

□

We recall the following result:

Theorem 9.2.21 (Gram–Schmidt Orthogonalization) *If $(V, \langle \cdot, \cdot \rangle)$ is an inner product space, and $\{v_1, v_2, v_3, \dots\}$ form a linearly independent set, then there is an orthonormal set $\{\phi_1, \phi_2, \phi_3, \dots\}$ such that for all n*

$$\text{span}\{v_1, \dots, v_n\} = \text{span}\{\phi_1, \dots, \phi_n\}$$

Proof: Suppose that $\{v_1, v_2, v_3, \dots\}$ is a linearly independent subset of V . We proceed by inductively building an orthonormal set $\{\phi_1, \phi_2, \phi_3, \dots\}$ so that

$$\text{span}\{v_1, \dots, v_n\} = \text{span}\{\phi_1, \dots, \phi_n\} \quad \text{for } i = 1, \dots, n$$

Define $\phi_1 := \frac{v_1}{\|v_1\|}$. Then $\|\phi_1\| = 1$, and certainly $\text{span}\{\phi_1\} = \text{span}\{v_1\}$.

Assume now that we have already defined ϕ_1, \dots, ϕ_n so that $\{\phi_1, \dots, \phi_n\}$ is an orthonormal set with the same span as $\{v_1, \dots, v_n\}$. We must now define ϕ_{n+1} . First define

$$w_{n+1} = v_{n+1} - \sum_{j=1}^n \langle v_{n+1}, \phi_j \rangle \phi_j$$

and note that

- (i) $\phi_{n+1} \neq 0$, for otherwise 0 would be a linear combination of ϕ_1, \dots, ϕ_n and v_n , and thus a linear combination of v_1, \dots, v_{n+1} . But v_1, \dots, v_{n+1} are linearly independent.
- (ii) If $1 \leq j \leq n$, then

$$\langle w_{n+1}, \phi_j \rangle = \langle v_{n+1}, \phi_j \rangle - \sum_{k=1}^n \langle v_{n+1}, \phi_k \rangle \langle \phi_k, \phi_j \rangle = 0$$

It follows that $\phi_1, \dots, \phi_n, w_{n+1}$ form an orthogonal set, and thus linearly independent.

- (iii) As $\text{span}\{\phi_1, \dots, \phi_n\} = \text{span}\{v_1, \dots, v_n\}$, we see that $\text{span}\{\phi_1, \dots, \phi_n, w_{n+1}\} = \text{span}\{v_1, \dots, v_{n+1}\}$.

The only potential problem is that we might not have $\|w_{n+1}\| = 1$. Therefore, define

$$\phi_{n+1} := \frac{w_{n+1}}{\|w_{n+1}\|}$$

□

9.3 Fourier Series

This section is adapted from *Measure Theory and Probability*, by Malcolm Adams and Victor Guillemin, Wadsworth 1986.

9.3.1 Statement of Results

We begin by stating the following theorem:

Theorem 9.3.1 *The functions*

$$\phi_n(x) := \frac{1}{\sqrt{2L}} e^{\frac{in\pi x}{L}} \quad n \in \mathbb{Z}$$

form a complete orthonormal basis for the Hilbert space $\mathcal{L}^2([-L, L], \mathcal{B}[-L, L], \lambda)$.

The fact that the ϕ_n form an orthonormal sequence is a rather straightforward exercise:

Exercise 9.3.2 Suppose that ϕ_n are defined as in Thm. 9.3.1. Show that $\langle \phi_n, \phi_m \rangle = \delta_{nm}$.

□

The fact that the ϕ_n form a complete set is more difficult, and left for the next subsection.

To say that the $(\phi_n)_{n \in \mathbb{Z}}$ form an orthonormal basis, rather than just an orthonormal set, means that every function $f \in \mathcal{L}^2[-L, L]$ can be represented as a series

$$f(x) = \sum_{n=-\infty}^{\infty} c_n e^{inx} \quad c_n := \frac{1}{2L} \int_{-L}^L f(x) e^{-in\pi x/L} dx$$

(By an expression such as $\sum_{n=-\infty}^{\infty} c_n e^{inx}$ we mean the limit $\lim_{N \rightarrow \infty} \sum_{n=-N}^N c_n e^{inx}$.) If we define $S_N(x) := \sum_{n=-N}^N c_n e^{in\pi x/L} = c_0 + \sum_{n=1}^N (c_n e^{in\pi x/L} + c_{-n} e^{-in\pi x/L})$, then this, in turn, means that $S_N \rightarrow f$ in \mathcal{L}^2 as $N \rightarrow \infty$.

Now note that

$$e^{in\pi(x+2L)/L} = e^{in\pi x/L} \quad \text{for all } n \in \mathbb{Z}$$

as $e^{2\pi i n} = 1$. It follows that if we regard the ϕ_n as functions on \mathbb{R} rather than just $[-L, L]$, then they are *periodic* with period $2L$ — a concept we now define:

Definition 9.3.3 A measurable function $f : (\mathbb{R}, \mathcal{B}(\mathbb{R})) \rightarrow \mathbb{C}$ is *periodic* with period a iff

$$f(x+a) = f(x) \quad \text{all } x \in \mathbb{R}$$

Note that any function $f : (-L, L] \rightarrow \mathbb{R}$ can be extended to a periodic function (with period $2L$) in a unique way. For if $x \in \mathbb{R}$, then there is $n \in \mathbb{Z}$ such that $(2n-1)L < x \leq (2n+1)L$. Then $-L < x - 2nL \leq L$, and necessarily we must have $f(x) := f(x - 2nL)$.

In many cases, however, we can get even stronger convergence the \mathcal{L}^2 -convergence:

Theorem 9.3.4 Suppose that f is continuous and periodic of period $2L$, and that it is piecewise differentiable on $[-L, L]$. Also let $c_n := \frac{1}{2L} \int_{-L}^L f(x) e^{-in\pi x/L} dx$. Then

$$\sum_{n=-\infty}^{\infty} c_n e^{in\pi x/L} \quad \text{converges to } f(x) \quad \text{uniformly in } x$$

i.e. for every $\varepsilon > 0$ there is $M \in \mathbb{N}$ such that

$$\left| \sum_{n=-N}^N c_n e^{in\pi x/L} - f(x) \right| < \varepsilon \quad \text{for all } x \text{ whenever } N \geq M$$

Although we find it simpler to prove results using the orthonormal basis $\frac{1}{\sqrt{2L}} e^{\frac{in\pi x}{L}}$, it is customary and useful to state these results in terms of sines and cosines. The next exercise shows how to go from one representation to the other, and thus that they are equivalent.

Exercise 9.3.5 (a) Note that

$$\cos \theta = \frac{e^{i\theta} + e^{-i\theta}}{2} \quad \sin \theta = \frac{e^{i\theta} - e^{-i\theta}}{2i}$$

Show that

$$\left\{ \frac{1}{\sqrt{2L}}, \frac{1}{\sqrt{L}} \cos(\pi x/L), \frac{1}{\sqrt{L}} \cos(2\pi x/L), \frac{1}{\sqrt{L}} \cos(3\pi x/L), \dots, \frac{1}{\sqrt{L}} \sin(\pi x/L), \frac{1}{\sqrt{L}} \sin(2\pi x/L), \frac{1}{\sqrt{L}} \sin(3\pi x/L) \dots \right\}$$

forms an orthonormal set in $\mathcal{L}^2[-L, L]$.

(b) Show that this orthonormal set is complete (i.e. forms an orthonormal basis) if and only if $\{\frac{1}{\sqrt{2L}}e^{inx} : n \in \mathbb{Z}\}$ forms an orthonormal basis.

(c) Show that if $n \in \mathbb{N}$, then

$$c_n e^{in\pi x/L} + c_{-n} e^{-in\pi x/L} = a_n \cos(n\pi x/L) + b_n \sin(n\pi x/L)$$

where

$$a_n = (c_n + c_{-n}) \quad b_n = i(c_n - c_{-n})$$

(d) Deduce that if

$$f(x) = \sum_{n=-\infty}^{\infty} c_n e^{inx} \quad c_n = \frac{1}{2L} \int_{-L}^L f(x) e^{-in\pi x/L} dx$$

then

$$f(x) = a_0/2 + \sum_{n=1}^{\infty} \left(a_n \cos(n\pi x/L) + b_n \sin(n\pi x/L) \right)$$

where

$$a_n = \frac{1}{L} \int_{-L}^L f(x) \cos \frac{n\pi x}{L} dx \quad b_n = \frac{1}{L} \int_{-L}^L f(x) \sin \frac{n\pi x}{L} dx$$

and vice versa.

□

9.3.2 Examples

It is useful to note that

$$\int_{-L}^L e^{in\pi x/L} dx = 0 \quad \text{for all } n \in \mathbb{Z}, n \neq 0$$

Prove it!

Exercise 9.3.6 Consider the saw-tooth function, namely the periodic extension of $f : [-L, L] \rightarrow \mathbb{R} : x \mapsto x$ to all of \mathbb{R} .

(a) Draw a graph of f .

(b) Show that the Fourier coefficients $c_n := \frac{1}{2L} \int_{-L}^L f(x) e^{-in\pi x/L} dx$ of f are given by

$$c_n = \frac{iL}{n\pi} (-1)^n$$

(c) Determine now the sine/cosine coefficients $a_n = \frac{1}{L} \int_{-L}^L f(x) \cos(n\pi x/L) dx$ and $b_n = \frac{1}{L} \int_{-L}^L f(x) \sin(n\pi x/L) dx$. Show that

$$a_n = 0 \quad \text{all } n \in \mathbb{Z} \quad b_n = \frac{2L}{n\pi} (-1)^{n+1}$$

- (d) Explain how you could have known in advance that the a_n are zero. [Hint: Think in terms of even and odd functions.]
- (e) Draw the graphs of $\sum_{n=1}^N a_n \sin(n\pi x/L)$ for $N = 1, 2, \dots, 6$ (Use, e.g., Excel, R or Matlab) to see what is going on.

Exercise 9.3.7 Consider the periodic extension of $f : [-L, L] \rightarrow \mathbb{R} : x \mapsto |x|$ to \mathbb{R}

- (a) Draw a graph of f .
- (b) Determine now the sine/cosine coefficients $a_n = \frac{1}{L} \int_{-L}^L f(x) \cos(n\pi x/L) dx$ and $b_n = \frac{1}{L} \int_{-L}^L f(x) \sin(n\pi x/L) dx$. Explain how you can be certain in advance that the b_n are zero [Hint: Think in terms of even and odd functions.]
- (c) Draw graphs of $\frac{a_0}{2} + \sum_{n=1}^N a_n \cos(n\pi x/L)$ for $N = 1, 2, \dots, 6$ (Use, e.g., Excel, R or Matlab) to see what is going on.

□

Exercise 9.3.8 Consider the periodic extension of $f : [-1, 1] \rightarrow \mathbb{R}$ to all of \mathbb{R} , defined by

$$f(x) := \begin{cases} -x & \text{if } -1 \leq x \leq 0 \\ x^2 & \text{if } 0 \leq x \leq 1 \end{cases}$$

- (a) Draw a graph of f .
- (b) Determine the Fourier coefficients a_n, b_n of f
- (c) Draw graphs of $\frac{a_0}{2} + \sum_{n=1}^N (a_n \cos(n\pi x) + b_n \sin(n\pi x))$ for $N = 1, 2, \dots, 6$ (Use, e.g., Excel, R or Matlab) to see what is going on.

□

9.3.3 Proofs*

For simplicity (no $\frac{\pi}{L}$'s) we restrict ourself to the case $L = \pi$. Throughout this section, therefore, let f be a periodic function with period 2π .

Note that if I is a closed interval of length 2π , then

$$\int_I f(x) dx = \int_{-\pi}^{\pi} f(x) dx \quad (\dagger)$$

For if $I = [a - \pi, a + \pi]$, then

$$\begin{aligned} \int_{a-\pi}^{a+\pi} f(x) dx &= \int_{-\pi}^{\pi} f(x) dx + \int_{\pi}^{a+\pi} f(x) dx - \int_{-\pi}^{a-\pi} f(x) dx \\ &= \int_{-\pi}^{\pi} f(x) dx + \int_{\pi}^{a+\pi} f(x) dx - \int_{\pi}^{a+\pi} f(x-2\pi) dx \\ &= \int_{-\pi}^{\pi} f(x) dx \end{aligned}$$

Define $c_n := \langle f, \phi_n \rangle = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} f(x) e^{-inx} dx$.

Theorem 9.3.9 Suppose that f is a continuous periodic function, with period 2π and that $x_0 \in [-\pi, \pi]$. Suppose further that the left- and right derivatives $\lim_{x \uparrow x_0} \frac{f(x) - f(x_0)}{x - x_0}$ and $\lim_{x \downarrow x_0} \frac{f(x) - f(x_0)}{x - x_0}$ exist at x_0 . Then

$$\frac{1}{2\pi} \sum_{n=-\infty}^{\infty} c_n e^{inx_0} = f(x_0)$$

(i.e. the series converges, and its limit is $f(x_0)$.)

The proof needs a small lemma:

Lemma 9.3.10 Define $D_N(x) := \frac{1}{2\pi} \sum_{n=-N}^N e^{inx}$. Then

$$(i) \int_{-\pi}^{\pi} D_N(x) dx = 1$$

$$(ii) D_N(x) = \frac{e^{i(N+1)x} - e^{-iNx}}{2\pi(e^{ix} - 1)}$$

Proof: To obtain (i), just observe that

$$\int_{-\pi}^{\pi} e^{inx} dx = \begin{cases} 2\pi & \text{if } n = 0 \\ 0 & \text{else} \end{cases}$$

(ii): Define $\alpha = e^{ix}$. Then

$$D_N(x) = \frac{1}{2\pi} \sum_{n=-N}^N \alpha^n = \frac{1}{2\pi} \alpha^{-N} \sum_{n=0}^{2N} e^{inx} = \frac{1}{2\pi} \alpha^{-N} \frac{\alpha^{2N+1} - 1}{\alpha - 1}$$

+

Proof of Thm. 9.3.9: Let $S_N := \frac{1}{\sqrt{2\pi}} \sum_{n=-M}^N c_n e^{inx_0} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \sum_{n=-M}^N e^{in(x_0-x)} dx$, and let $D_N(x) := \frac{1}{2\pi} \sum_{n=-N}^N e^{inx}$ so that, using a change of variables and (†), we have

$$S_N = \int_{-\pi}^{\pi} f(x) D_N(x_0 - x) dx = \int_{-\pi}^{\pi} f(x_0 - x) D_N(x) dx$$

According to (i) of the preceding lemma, we have $\int_{-\pi}^{\pi} D_N(x) dx = 1$, so that $\int_{-\pi}^{\pi} f(x_0) D_N(x) dx = f(x_0)$. Hence

$$S_N - f(x_0) = \int_{-\pi}^{\pi} [f(x_0 - x) - f(x_0)] D_N(x) dx$$

Define now

$$g(x) := \frac{f(x_0 - x) - f(x_0)}{e^{ix} - 1} = \frac{f(x_0 - x) - f(x_0)}{x} \left(\frac{x}{e^{ix} - 1} \right)$$

Now consider the two factors on the right in the above equation: The first factor $\frac{f(x_0-x)-f(x_0)}{x}$ is defined and continuous everywhere on $[-\pi, \pi]$ except at $x = 0$. Nevertheless $\lim_{x \uparrow 0} \frac{f(x_0-x)-f(x_0)}{x}$ and $\lim_{x \downarrow 0} \frac{f(x_0-x)-f(x_0)}{x}$ exist by assumption. The second factor $\frac{x}{e^{ix}-1}$ is defined and continuous everywhere on $[-\pi, \pi]$ except at $x = 0$. Nevertheless, according to L'Hôpital's Rule,

$\lim_{x \rightarrow 0} \frac{x}{e^{ix} - 1} = -i$. It is clear therefore, that $g(x)$ is defined and continuous everywhere on $[-\pi, \pi]$ except at $x = 0$, but that $g(0-) := \lim_{x \uparrow} g(x)$ and $g(0+) := \lim_{x \downarrow} g(x)$ exist. It is therefore clear that $g \in \mathcal{L}^2([-\pi, \pi])$.

Using (ii) of the preceding lemma, we have

$$S_N - f(x_0) = \int_{-\pi}^{\pi} [f(x_0 - x) - f(x_0)] D_N(x) dx = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(x) e^{i(N+1)x} dx - \frac{1}{2\pi} \int_{-\pi}^{\pi} g(x) e^{-iNx} dx$$

and hence

$$S_N - f(x_0) = \frac{1}{\sqrt{2\pi}} (d_{-(N+1)} - d_N)$$

where d_n is the n^{th} Fourier coefficient of f . As $d_n \rightarrow 0$ as $n \rightarrow \pm\infty$ by Corollary 9.2.15, we see that $S_N \rightarrow f(x_0)$ as $N \rightarrow \infty$.

—

A function f is said to be piecewise differentiable on a compact interval if and only if it is the derivative $f'(x)$ exists and is continuous every where except at possibly finitely many points. For such functions, we get much better convergence. Before we state this result, recall this simple but useful result for obtaining uniform convergence from pointwise convergence:

Exercise 9.3.11 Suppose that $a_n(x)$ are functions and that $\sum_{n=1}^{\infty} a_n(x) = f(x)$ for all x i.e. the series converges, for all x to $f(x)$. Suppose further that there are real numbers r_n such that $|a_n(x)| < r_n$ for all x and all n . Show that if $\sum_n r_n < \infty$, then $\sum_{n=1}^{\infty} a_n$ converges uniformly to f , i.e. show that in that case, for every $\varepsilon > 0$ we can find N such that for all $n \geq N$ and all x we have $|\sum_{k=1}^n a_k(x) - f(x)| < \varepsilon$.

[Hint: Let $\varepsilon > 0$, and choose N so that $\sum_{k=n}^{\infty} r_k < \varepsilon$ for all $n \geq N$. (Why can we do this?) Now note $|\sum_{k=1}^n a_k(x) - \sum_{k=1}^m a_k(x)| \leq \sum_{k=n+1}^m r_k \leq \sum_{k=n+1}^{\infty} r_k < \varepsilon$ for all $n \geq M$, and all x . Let $m \rightarrow \infty$ to get $|\sum_{k=1}^n a_k(x) - f(x)| < \varepsilon$ for all $n \geq N$ and all x .]

□

Next, note the following:

Lemma 9.3.12 Suppose that f is continuous and periodic of period 2π , and that it is piecewise differentiable on $[-\pi, \pi]$. For $x \in [-\pi, \pi]$ let $g(x) := f'(x)$ where $f'(x)$ is defined, and let $g(x)$ be arbitrary (e.g. set $g(x) := 0$) otherwise. Let c_n, d_n be the Fourier coefficients of f, g respectively. Then

$$d_n = inc_n$$

Proof: As f is piecewise differentiable, there are $\pi - a_0 < a_1 < \cdots < a_p = \pi$ so that g is continuous on (a_i, a_{i+1}) . Then, integrating by parts

$$\begin{aligned} \sqrt{2\pi} d_n &= \int_{-\pi}^{\pi} g(x) e^{-inx} dx = \sum_{i=1}^p \int_{a_{i-1}}^{a_i} g(x) e^{-inx} dx \\ &= \sum_{i=1}^p \left[f(x) e^{-inx} \Big|_{a_{i-1}}^{a_i} + in \int_{a_{i-1}}^{a_i} f(x) e^{-inx} dx \right] \\ &= f(x) e^{-inx} \Big|_{-\pi}^{\pi} + in \int_{-\pi}^{\pi} f(x) e^{-inx} dx \\ &= in \int_{-\pi}^{\pi} f(x) e^{-inx} dx = \sqrt{2\pi} inc_n \end{aligned}$$

(because $f(\pi) e^{in\pi} = f(-\pi) e^{-in\pi}$ by periodicity of f).

Theorem 9.3.13 Suppose that f is continuous and periodic of period 2π , and that it is piecewise differentiable on $[-\pi, \pi]$. Also let $c_n := \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} f(x) e^{-inx} dx$. Then

$$\frac{1}{\sqrt{2\pi}} \sum_{n=-\infty}^{\infty} c_n e^{inx} \quad \text{converges to } f(x) \quad \text{uniformly in } x$$

i.e. for every $\varepsilon > 0$ there is $M \in \mathbb{N}$ such that

$$\left| \frac{1}{\sqrt{2\pi}} \sum_{n=-N}^N c_n e^{inx} - f(x) \right| < \varepsilon \quad \text{for all } x \text{ whenever } N \geq M$$

Proof of Thm. 9.3.13: For each x , let $S_N(x) := \frac{1}{\sqrt{2\pi}} \sum_{n=-N}^N c_n e^{inx}$. Then by Thm. 9.3.9, we have that $S_N(x) \rightarrow f(x)$ for each x . Note that $|c_n e^{inx}| = |c_n|$, so if we can show that $\sum_{n=-\infty}^{\infty} |c_n| < \infty$, then the preceding exercise will yield that S_N converges to f uniformly. Now let d_n be the Fourier coefficients of $g = f'$. Since g is piecewise continuous, $g \in \mathcal{L}^2$, and hence $\sum_{n=-\infty}^{\infty} |d_n|^2 = \|g\|^2 < \infty$. By Lemma 9.3.12, we see that $\sum_{n=-\infty}^{\infty} n^2 |c_n|^2 < \infty$. The Cauchy-Schwarz inequality for counting measure on \mathbb{Z} yields that

$$\sum_n |c_n| \leq |c_0| + \sum_{n \neq 0} \left(\frac{1}{|n|} \right) (|n| |c_n|) \leq |c_0| + \left(\sum_{n \neq 0} \frac{1}{n^2} \right) \left(\sum_{n \neq 0} |n|^2 |c_n|^2 \right) < \infty$$

+

At this point, it remains to show that the sequence $\phi_n(x) := \frac{1}{\sqrt{2\pi}} e^{inx}$ is an orthonormal basis. In order to do this, it suffices to show that

$$\langle f, \phi_n \rangle = 0 \quad \text{for all } n \in \mathbb{Z} \quad \text{implies} \quad f = 0 \text{ a.e.}$$

We first note that:

Lemma 9.3.14 If $\langle f, \phi_n \rangle = 0$ for all $n \in \mathbb{Z}$, then

$$\int_a^b f(x) dx = 0 \quad \text{for all subintervals } [a, b] \subseteq [-\pi, \pi]$$

Proof: For $\varepsilon > 0$, define on $[-\pi, \pi]$ the function

$$J_\varepsilon(x) := \frac{x-a}{\varepsilon} I_{[a, a+\varepsilon)}(x) + I_{[a+\varepsilon, b-\varepsilon]}(x) + \frac{b-x}{\varepsilon} I_{(b-\varepsilon, b]}$$

Note that J_ε is a continuous piecewise differentiable approximation of the indicator function $I_{[a, b]}$, with $J_\varepsilon \uparrow I_{[a, b]}$ as $\varepsilon \downarrow 0$.

It follows from Thm. 9.3.13 that $S_N(J_\varepsilon) \rightarrow J_\varepsilon$ uniformly (where $(S_N(J_\varepsilon))_N$ is the sequence of partial Fourier sums for the function J_ε). It follows easily that also $S_N(J_\varepsilon) \rightarrow J_\varepsilon$ in \mathcal{L}^2 , and thus that

$$\langle f, S_N(J_\varepsilon) \rangle \rightarrow \langle f, J_\varepsilon \rangle$$

But $\langle f, \phi_n \rangle = 0$ for all n , and hence $\langle f, S_N(J_\varepsilon) \rangle = 0$ for all N , from which it follows that $\langle f, J_\varepsilon \rangle = 0$. Thus by MCT

$$\int_a^b f \, dx = \int_{-\pi}^{\pi} f I_{[a,b]} \, dx = \lim_{\varepsilon \downarrow 0} \int_{-\pi}^{\pi} f J_\varepsilon \, dx = \lim_{\varepsilon \downarrow 0} \langle f, J_\varepsilon \rangle = 0$$

+

Proof of Thm. 9.3.1: Suppose that $\langle f, \phi_n \rangle = 0$ for all $n \in \mathbb{Z}$. Let

$$\mathcal{D} := \left\{ B \in \mathcal{B}[-\pi, \pi] : \int_B f \, dx = 0 \right\}$$

It is easy to see that \mathcal{D} is a λ -system. By the previous lemma, \mathcal{D} contains all closed subintervals of $[-\pi, \pi]$, and hence, by the Monotone Class Theorem ($\pi - \lambda$ version) we have $\mathcal{D} = \mathcal{B}[-\pi, \pi]$. It follows immediately that $f = 0$ a.e.

+

Chapter 10

Weak Convergence and Characteristic Functions

10.1 Weak Convergence and Convergence in Distribution

We will scratch only the surface of the theory of *weak convergence*, also known as *convergence in distribution* of probability measures, and restrict ourselves entirely to measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Define

$$\mathcal{M}_1(\mathbb{R}) := \text{set of all probability measures on } \mathbb{R}$$

and

$$\mathcal{C}_b(\mathbb{R}) := \text{set of all bounded continuous functions } \mathbb{R} \rightarrow \mathbb{R}$$

Recall also that if $X : (\Omega, \mathcal{F}, \mathbb{P}) \rightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is a random variable, then the *law* (or *distribution*) of X is a probability measure μ_X on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ defined by

$$\mu_X(B) := \mathbb{P}X^{-1}(B) = \mathbb{P}(X \in B)$$

The definition of weak convergence uses both measure-theoretic and topological properties:

Definition 10.1.1 (a) Let $\mu_n, \mu \in \mathcal{M}_1(\mathbb{R})$ for $n \in \mathbb{N}$. We say that $(\mu_n)_n$ converges weakly to μ , and write

$$\mu_n \xrightarrow{w} \mu$$

iff for each $f \in \mathcal{C}_b(\mathbb{R})$ we have

$$\lim_{n \rightarrow \infty} \int f d\mu_n = \int f d\mu$$

(b) If X_n, X are real-valued random variables, we say that $(X_n)_n$ converges in distribution to X if and only if $\mu_{X_n} \xrightarrow{w} \mu_X$.

Remarks 10.1.2 Note that, in the definition of convergence in distribution of random variables, the X_n, X are not required to be defined on the same space — it is only their laws that matter. Nevertheless, if the X_n, X are all defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then, seeing that $\int f d\mu_X = \mathbb{E}[f(X)]$ by the change of variables formula for integrals, we have

$$X_n \xrightarrow{w} X \quad \text{iff} \quad \mathbb{E}[f(X_n)] \rightarrow \mathbb{E}[f(X)] \quad \text{for all } f \in \mathcal{C}_b(\mathbb{R})$$

□

Exercise 10.1.3 1. Show that if $X_n \xrightarrow{\text{a.s.}} X$, then $X_n \xrightarrow{w} X$.

2. Show that if $X_n \xrightarrow{\mathcal{L}^1} X$, then $X_n \xrightarrow{w} X$.

□

In elementary (non-measure-theoretic) probability texts, the definition of convergence in distribution is often given in terms of *distribution functions*. Recall that a function $F : \mathbb{R} \rightarrow [0, 1]$ is a distribution function provided that

(i) F is increasing, i.e. $x \leq y \Rightarrow F(x) \leq F(y)$.

(ii) $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow +\infty} F(x) = 1$.

(iii) F is right-continuous, i.e. $\lim_{y \downarrow x} F(y) = F(x)$ for all $x \in \mathbb{R}$.

As F is increasing, $F(x-) := \lim_{y \uparrow x} F(y)$ exists. F is discontinuous at x precisely when $F(x) \neq F(x-)$. We shall call a point $x \in \mathbb{R}$ where a distribution function F is discontinuous an **atom** of the distribution function. If F is the distribution function of some random variable x , then x is an atom of F if and only if $\mathbb{P}(X = x) = F(x) - F(x-) > 0$.

Recall also that every probability measure μ on \mathbb{R} induces a distribution function: Simply define $F(x) = \mu(-\infty, x]$. The converse is also true: Any distribution function F yields a unique probability measure μ on $(\mathbb{R}, \mathcal{B})$ defined by $\mu(-\infty, x] = F(x)$, and which can then be extended to all Borel sets by Carathéodory's extension theorem.

Definition 10.1.4 (Non-measure-theoretic definition of convergence in distribution)

Let F_{X_n}, F_X be the distribution functions of random variables X_n, X .

We say that X_n converges to X in distribution, written $X_n \xrightarrow{d} X$, provided that

$$F_{X_n}(x) \rightarrow F_X(x) \text{ as } n \rightarrow +\infty$$

at every point x where F_X is continuous.

Example 10.1.5 Consider constant random variables $X_n := \frac{1}{n}$ on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The law of X_n is just the Dirac delta $\delta_{\frac{1}{n}}$, i.e. the point mass at $\frac{1}{n}$. Clearly, we ought to have $\delta_{\frac{1}{n}} \rightarrow \delta_0$, as the points $\frac{1}{n}$ lie closer and closer to the point 0. Nevertheless, we do not have $\delta_{\frac{1}{n}}(B) \rightarrow \delta_0(B)$ for every Borel set: For example

$$\delta_{\frac{1}{n}}\{0\} \not\rightarrow \delta_0\{0\}$$

as $0 \neq 1$. Similarly, we do not have

$$\int f(x) \delta_{\frac{1}{n}}(dx) \rightarrow \int f(x) \delta_0(dx) \quad \text{for all measurable functions}$$

take $f = I_{\{0\}}$ or $f = I_{(-\infty, 0]}$, for example.

However, the notion of $\frac{1}{n}$ lying “closer and closer” to 0 is a topological notion. If f respects topology, i.e. if f preserves the “closeness”-relation, i.e. if f is continuous, then we have $\lim_n f(\frac{1}{n}) = f(0)$, i.e.

$$\lim_n \int f(x) \delta_{\frac{1}{n}}(dx) = \lim_n f(\frac{1}{n}) = f(0) = \int f(x) \delta_0(dx)$$

□

Lemma 10.1.6 *A distribution function can have at most countably many points where it is discontinuous.*

Proof: Let $D := \{x \in \mathbb{R} : F(x) - F(x-) > 0\}$ be the set of all points where F is discontinuous, and, for $n \in \mathbb{N}$, define $D_n = \{x \in \mathbb{R} : F(x) - F(x-) > \frac{1}{n}\}$. Clearly, $D = \bigcup_n D_n$. But since $0 \leq F(x) \leq 1$ for all x , and since F is increasing, each D_n can have at most n elements.

—

Proposition 10.1.7 *Let F_n, F be distribution functions on \mathbb{R} for $n \in \mathbb{N}$ (in the sense defined above), and that $F_n(x) \rightarrow F(x)$ at every point x where F is continuous. Then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ carrying random variables X_n, X (for $n \in \mathbb{N}$) such that*

$$F_n = F_{X_n}, \quad F = F_X \quad \text{and} \quad X_n \xrightarrow{\text{a.s.}} X$$

Proof: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be $([0, 1], \mathcal{B}[0, 1], \lambda)$ and define

$$\begin{aligned} X^+(\omega) &= \inf\{y \in \mathbb{R} : F(y) > \omega\} = \sup\{y \in \mathbb{R} : F(y) \leq \omega\} \\ X^-(\omega) &= \inf\{y \in \mathbb{R} : F(y) \geq \omega\} = \sup\{y \in \mathbb{R} : F(y) < \omega\} \end{aligned}$$

and define $X_n^+(\omega)$ and $X_n^-(\omega)$ similarly.

Note that we always have $X^-(\omega) \leq X^+(\omega)$. If $\omega \in [0, 1]$, then $[X^-(\omega), X^+(\omega)]$ is a closed subinterval of \mathbb{R} , possibly degenerate, i.e. consisting of just a single point. When the interval $[X^-(\omega), X^+(\omega)]$ is non-degenerate, then F is constant with value ω on that interval. Clearly, therefore, if $[X^-(\omega_1), X^+(\omega_1)] = [X^-(\omega_2), X^+(\omega_2)]$ is non-degenerate, then $\omega_1 = \omega_2$. As each non-degenerate interval must contain a rational number, there are at most countably many ω for which $[X^-(\omega), X^+(\omega)]$ is non-degenerate, i.e. at most countably many $\omega \in [0, 1]$ for which $X^+(\omega) \neq X^-(\omega)$. Thus $X^+ = X^-$ λ -a.s.

Note that by right-continuity of F , we have

$$\omega \leq F(x) \quad \Longleftrightarrow \quad X^-(\omega) \leq x$$

and thus

$$\mathbb{P}(X^- \leq x) = \lambda\{\omega \in [0, 1] : X^-(\omega) \leq x\} = \lambda[0, F(x)] = F(x)$$

This shows that the distribution function of X^- is F . As $X^+ = X^-$ a.s., X^+ has the same distribution as X^- , i.e. the distribution function of X^+ is also F .

In the same way it follows that the random variables X_n^+ and X_n^- have distribution functions F_n .

For definiteness, define $X := X^+$. We now show that $X_n^+ \xrightarrow{\text{a.s.}} X$ and that $X_n^- \xrightarrow{\text{a.s.}} X$. For let $\omega \in \Omega$, and let y be a point of continuity of F such that $y > X^+(\omega)$. Then $F(y) > \omega$, and hence, for all sufficiently large n , also $F_n(y) > \omega$ (because $F_n(y) \rightarrow F(y)$). Hence $y \geq X_n^+(\omega)$ for all sufficiently large n , so that $\limsup X_n^+(\omega) \leq y$. Now since there are at most countably many points where F is not continuous, we must have

$$\limsup X_n^+(\omega) \leq X^+(\omega)$$

In the same way, we can prove that

$$\liminf X_n^-(\omega) \geq X^-(\omega)$$

Putting these inequalities together, we see that

$$X^-(\omega) \leq \liminf X_n^-(\omega) \leq \limsup X_n^+(\omega) \leq X^+(\omega)$$

Now since $X^- = X^+$ a.s., we must have $X_n^- \xrightarrow{\text{a.s.}} X$ and $X_n^+ \xrightarrow{\text{a.s.}} X$.

+

Theorem 10.1.8 Suppose that μ_n, μ are probability distributions on \mathbb{R} and that F_n, F are the associated distribution functions. Then $\mu_n \xrightarrow{w} \mu$ if and only if $F_n(x) \rightarrow F(x)$ at every point $x \in \mathbb{R}$ where F is continuous.

Proof: First suppose that $\mu_n \xrightarrow{w} \mu$. Let $x \in \mathbb{R}$ and let $\delta > 0$. Define a bounded continuous function f by

$$f(y) = \begin{cases} 1 & \text{if } y \leq x \\ 1 - \delta^{-1}(y - x) & \text{if } x < y < x + \delta \\ 0 & \text{if } y \geq x + \delta \end{cases}$$

Thus if δ is small, then f is a continuous approximation of a step function which jumps from 1 to 0 at x .

Note that $I_{(-\infty, x]} \leq f \leq I_{(-\infty, x+\delta]}$. Since $\mu_n \xrightarrow{w} \mu$, we must have

$$\limsup_n F_n(x) = \limsup_n \int I_{(-\infty, x]} d\mu_n \leq \limsup \int f d\mu_n = \int f d\mu \leq \int I_{(-\infty, x+\delta]} d\mu = F(x+\delta)$$

Using right continuity of F , we see, upon letting $\delta \downarrow 0$, that $\limsup_n F_n(x) \leq F(x)$ for all $x \in \mathbb{R}$.

Similarly, define

$$g(y) = \begin{cases} 1 & \text{if } y \leq x - \delta \\ 1 - \delta^{-1}(y - (x - \delta)) & \text{if } x - \delta < y < x \\ 0 & \text{if } y \geq x \end{cases}$$

so that $I_{(-\infty, x-\delta]} \leq g \leq I_{(-\infty, x]}$. Then

$$F_n(x-) = \mu_n(-\infty, x) \geq \int g d\mu_n$$

Since $\mu_n \xrightarrow{w} \mu$, we must have

$$\liminf_n F_n(x-) \geq \liminf_n \int g d\mu_n = \int g d\mu \geq F(x - \delta)$$

Letting $\delta \downarrow 0$, we get

$$\liminf_n F_n(x-) \geq F(x-)$$

for all $x \in \mathbb{R}$.

It now follows that

$$F(x-) \leq \liminf_n F_n(x-) \leq \limsup_n F_n(x) \leq F(x)$$

In particular, if x is a point of continuity of F (i.e. if $F(x-) = F(x)$), then

$$\lim_n F_n(x) = F(x)$$

as required. This proves the forward direction.

Now assume that $F_n(x)$ converges to $F(x)$ at every continuity point of F . There is a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ which carries random variables X_n, X with the properties that

$$F_{X_n} = F_n \quad F_X = F \quad \text{and} \quad X_n \xrightarrow{\text{a.s.}} X$$

If $f \in \mathcal{C}_b(\mathbb{R})$, bounded by a constant K , then the $f(X_n)$ are random variables which are bounded by K as well. By the Lebesgue Dominated Convergence Theorem, we therefore have

$$\int f d\mu_n = \int f(X_n) d\mathbb{P} \rightarrow \int f(X) d\mathbb{P} = \int f d\mu$$

Thus $\mu_n \xrightarrow{w} \mu$, proving the reverse direction.

□

We now seek a kind of compactness condition for probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Definition 10.1.9 A sequence $(\mu_n)_n$ of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is said to be *tight* if and only if

$$\sup_n \mu_n\{x : |x| > K\} \rightarrow 0 \quad \text{as} \quad K \rightarrow \infty$$

A sequence of distribution functions (F_n) on \mathbb{R} is tight if and only if the corresponding probability distributions form a tight sequence.

Remarks 10.1.10 (a) You should verify the following directly from the definition: $(\mu_n)_n$ is tight if and only if for every $\varepsilon > 0$ there exists a $K > 0$ such that

$$\mu_n[-K, K] > 1 - \varepsilon \quad \text{for all } n \in \mathbb{N}$$

In other words, most of the mass of each μ_n lies on a single compact interval $[-K, K]$, which is the same for all μ_n .

(b) It should be clear that a single probability distribution μ on \mathbb{R} is tight, i.e. that, for any $\varepsilon > 0$ there is K such that $\mu\{x : |x| > K\} < \varepsilon$. Indeed, since $[-n, n] \uparrow \mathbb{R}$, we have $\mu[-n, n] \uparrow 1$, and thus there is K such that $\mu[-K, K] > 1 - \varepsilon$.

In the same way, it can be shown that any finite set $\{\mu_1, \dots, \mu_n\}$ is tight: Choose K_j so that $\mu_j[-K_j, K_j] > 1 - \varepsilon$, and then define $K := \max\{K_1, \dots, K_n\}$. Clearly $\mu_j[-K, K] \geq \mu_j[-K_j, K_j] > 1 - \varepsilon$ for all $j = 1, \dots, n$.

(c) If μ_n is the distribution of a $N(m_n, 1)$ -normal random variable, and the sequence of means (m_n) is bounded, then $(\mu_n)_n$ is tight.

(d) If μ_n is the distribution of a $N(0, n)$ -normal random variable, then $(\mu_n)_n$ is not tight.

(e) If $\mu_n = \delta_n$, then $(\mu_n)_n$ is not tight.

□

Theorem 10.1.11 (Helly–Bray)

(a) Let F_n be a sequence of distribution functions on \mathbb{R} . Then there exists a right-continuous non-decreasing function $F : \mathbb{R} \rightarrow [0, 1]$ and a subsequence F_{n_k} such that

$$\lim_{k \rightarrow \infty} F_{n_k}(x) = F(x)$$

at every point of continuity of F .

(b) If, moreover, the $(F_n)_n$ are tight, then F is a distribution function, and $F_n \xrightarrow{w} F$.

Proof: (a) Enumerate the rationals (or any other countable dense subset of \mathbb{R}), i.e. $\mathbb{Q} = \{q_n : n \in \mathbb{N}\}$. Since the sequence

$$(F_n(q_1))_n \quad n \in \mathbb{N}$$

is bounded (because it lies in $[0, 1]$) it must have a convergent subsequence, by the Bolzano–Weierstrass theorem, i.e. there exists a sequence $(n_k^{(1)})_k$ in \mathbb{N} such that $F_{n_k^{(1)}}(q_1)$ converges to some value $\bar{F}(q_1)$ as $k \rightarrow +\infty$. Since $F_{n_k^{(1)}}(q_2)$ is bounded, it has a convergent subsequence, i.e. there exists a subsequence $(n_k^{(2)})_k$ of $(n_k^{(1)})_k$ such that $F_{n_k^{(2)}}(q_2)$ converges to some value $\bar{F}(q_2)$ as $k \rightarrow +\infty$. Since a subsequence of a convergent sequence converges, and to the same value, we also have $F_{n_k^{(2)}}(q_1) \rightarrow \bar{F}(q_1)$. Keep going in this way: Since $F_{n_k^{(2)}}(q_3)$ is bounded, there is a subsequence $(n_k^{(3)})_k$ of $(n_k^{(2)})_k$ such that $F_{n_k^{(3)}}(q_3)$ converges to some value $\bar{F}(q_3)$. Then, since $(n_k^{(3)})_k$ is a subsequence of $(n_k^{(2)})_k$ and thus also of $(n_k^{(1)})_k$, we must also have $F_{n_k^{(3)}}(q_2) \rightarrow \bar{F}(q_2)$ and $F_{n_k^{(3)}}(q_1) \rightarrow \bar{F}(q_1)$.

We now pick the diagonal of the sequences $n_k^{(i)}$: Define $n_k := n_k^{(k)}$. Note that the k^{th} tail $\{n_m : m \geq k\}$ of $(n_k)_k$ is a subsequence of $(n_j^{(i)})_j$ for each $i \geq k$. It follows that

$$F_{n_k}(q_i) \rightarrow \bar{F}(q_i) \quad \text{for all } q_i \in \mathbb{Q}$$

Clearly \bar{F} is an increasing function $\mathbb{Q} \rightarrow [0, 1]$, but it hasn't been defined for all $x \in \mathbb{R}$. For every real number x , however, we can find a strictly decreasing sequence of rationals $q \downarrow x$. Thus define

$$F(x) = \lim_{q \downarrow x} \bar{F}(q)$$

where q strictly decreases to x .

Then F has all the required properties. (Note, however, that $F(q)$ and $\bar{F}(q)$ may not be equal). F is clearly increasing. To see that it is right-continuous, let $x \in \mathbb{R}$ and $\varepsilon > 0$. Choose $q \in \mathbb{Q}$ such that $x < q$, and $\bar{F}(q) < F(x) + \varepsilon$. This can be done by definition of F . Then

$$F(x) \leq \bar{F}(q) \leq F(x) + \varepsilon$$

which proves right-continuity.

Finally, suppose that F is continuous at x . If $\varepsilon > 0$, we may choose $y < x$ such that $F(x) - \varepsilon < F(y)$ (by continuity). We may also pick $q, r \in \mathbb{Q}$ such that $y < r < x < q$ and $\bar{F}(q) < F(x) + \varepsilon$. Since

$$F(x) - \varepsilon < \bar{F}(r) \leq \bar{F}(q) < F(x) + \varepsilon$$

and

$$F_n(r) \leq F_n(x) \leq F_n(q)$$

it follows that

$$F(x) - \varepsilon \leq \liminf_{k \rightarrow \infty} F_{n_k}(r) \leq \limsup_{k \rightarrow \infty} F_{n_k}(x) \leq F(x) + \varepsilon$$

Thus $F_{n_k}(x) \rightarrow F(x)$ at every point x where F is continuous.

(b) We need only show that the function F obtained in the Helly–Bray Lemma exhibits the correct end behaviour, since we already know that it is non-decreasing and right-continuous.

But this is easy: For example, to show that $\lim_{x \rightarrow +\infty} F(x) = 1$, proceed as follows. Let $\varepsilon > 0$. By tightness, we can find a $K > 0$ such that $F_n(K) - F_n(-K) > 1 - \varepsilon$. Moreover, K can be taken to be a point of continuity of F , because F has at most countably many atoms. Then $F_n(K) > 1 - \varepsilon$ for all n . Thus if $x > K$ is a point of continuity of F , then $F(x) \geq F(K) = \lim_{k \rightarrow +\infty} F_{n_k}(K) \geq 1 - \varepsilon$.

—

Translating from the language of distribution functions to that of distributions, we obtain

Corollary 10.1.12 *If $(\mu_n)_n$ is a tight sequence of probability distributions on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, then it has a weakly convergent subsequence, i.e. there is a probability distribution μ and a subsequence $(\mu_{n_k})_k$ so that $\mu_{n_k} \xrightarrow{w} \mu$.*

10.2 Characteristic Functions

10.2.1 Basic Properties

In this section we introduce the characteristic function of a random variable. This is intimately related to the *Fourier transform*, although this kinship will become apparent only later in this section.

Definition 10.2.1 (a) If μ is a probability measure on $(\mathbb{R}, \mathcal{B})$, then the **characteristic function** of μ is the function $\varphi : \mathbb{R} \rightarrow \mathbb{C}$ defined by

$$\varphi(t) = \int e^{itx} \mu(dx)$$

(b) If X is a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then the characteristic function $\mathbb{R} \xrightarrow{\varphi_X} \mathbb{C}$ of X is defined to be the characteristic function of the distribution μ_X of X .

Remarks 10.2.2 (a) Clearly identically distributed random variables have identical characteristic functions. We shall soon prove the converse to this, i.e. that random variables with identical characteristic functions are also identically distributed.

- (b) Suppose that X is a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. By the Change of Variables Theorem, we have

$$\varphi_X(t) = \int_{\mathbb{R}} e^{itx} d\mu_X = \int_{\Omega} e^{itX} d\mathbb{P} = \mathbb{E}[e^{itX}]$$

The characteristic function of a random variable X is therefore frequently defined by

$$\varphi_X(t) = \mathbb{E}[e^{itX}]$$

- (c) In a PDE's course, the Fourier transform $F = \mathcal{F}(f)$ of a piece-wise continuous real-valued function $f(x)$ is generally defined by

$$F(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx} f(x) dx$$

Thus, if X is a continuous random variable with density f_X , then the characteristic function of X is just the Fourier transform of f_X (apart from a factor of $\sqrt{2\pi}$). To be precise:

$$\varphi_X = \sqrt{2\pi} \mathcal{F}(f_X)$$

- (d) Of course,

$$\begin{aligned} \varphi_X(t) &= \int \cos(tX) d\mathbb{P} + i \int \sin(tX) d\mathbb{P} \\ &= \mathbb{E} \cos(tX) + i \mathbb{E} \sin(tX) \end{aligned}$$

□

Because the characteristic function of a continuous random variable is just the Fourier transform of its density function (up to a constant factor) the two have very similar behavioural properties. Note, however, that the characteristic function of a random variable X is defined even if X is not continuous: Since $|e^{itX}| = 1$, the random variable $e^{itX} = \cos tX + i \sin tX$ is bounded, and hence integrable.

Proposition 10.2.3 (Basic Properties of Characteristic Functions)

Let φ_X be the characteristic function of some random variable X .

- (a) $\varphi_X(0) = 1$
- (b) $|\varphi_X(t)| \leq 1$ for all $t \in \mathbb{R}$
- (c) φ_X is continuous (in fact, uniformly continuous)
- (d) $\varphi_{aX+b}(t) = e^{ibt} \varphi_X(at)$
- (e) $\varphi_{-X}(t) = \varphi_X(-t) = \overline{\varphi_X(t)}$ (= complex conjugate of $\varphi_X(t)$)

Proof: (a) is obvious.

(b) follows from the fact that $|\int f d\mu| \leq \int |f| d\mu$.

(c) Note that $e^{iuX} \rightarrow e^{itX}$ as $u \rightarrow t$. The result follows by the Lebesgue Dominated Convergence Theorem, as the family of e^{iuX} is dominated by an integrable random variable ($|e^{iuX}| = 1$). Uniform continuity is now easy to see.

(d) is straightforward.

(e) follows from the fact that both $\varphi_{-X}(t)$ and $\varphi_X(-t)$ are equal to $\mathbb{E}[e^{-itX}]$. Now

$$\mathbb{E}[e^{-itX}] = \mathbb{E} \cos(tX) - i \mathbb{E} \sin(tX) = \overline{\mathbb{E} \cos(tX) + i \mathbb{E} \sin(tX)} = \overline{\mathbb{E}[e^{itX}]}$$

□

The following trivial result is nevertheless of great importance:

Theorem 10.2.4 *If X, Y are independent random variables on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then*

$$\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$$

Exercise 10.2.5 Prove the preceding theorem. □

Examples 10.2.6 We give here some examples of characteristic functions of random variables:

- (a) **Normal distribution:** Assume that X is normally distributed with mean 0 and variance 1. Then it has density function

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Thus

$$\varphi_X(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} \cos tx \, dx + \frac{i}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} \sin tx \, dx$$

However, $e^{-x^2/2} \sin tx$ is an *odd* function of x , and thus the second integral on the right is 0. It follows that

$$\varphi_X(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} \cos tx \, dx$$

Differentiating with respect to t yields

$$\varphi'_X(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} -x e^{-\frac{x^2}{2}} \sin tx \, dx$$

If we integrate this integral by parts, we obtain

$$\begin{aligned} \varphi'_X(t) &= -\frac{t}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} \cos tx \, dx \\ &= -t\varphi_X(t) \end{aligned}$$

This is a first-order separable differential equation with solution

$$\varphi_X(t) = \varphi_X(0) e^{-\frac{t^2}{2}}$$

Since $\varphi_X(0) = 1$, we thus obtain

$$\varphi_X(t) = e^{-\frac{t^2}{2}}$$

Now if $Y = \sigma X + \mu$, then Y is a normally distributed random variable with mean μ and variance σ^2 . It follows that

$$\begin{aligned} \varphi_Y(t) &= \varphi_{\sigma X + \mu}(t) = e^{i\mu t} \varphi_X(\sigma t) \\ &= e^{i\mu t - \frac{1}{2}t^2\sigma^2} \end{aligned}$$

- (b) Suppose that X is a Bernoulli variable, with $\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = \frac{1}{2}$. Then

$$\varphi_X(t) = \mathbb{E}e^{itX} = \frac{e^{it}}{2} + \frac{e^{-it}}{2} = \cos t$$

- (c) Suppose that X is Poisson distributed with rate λ . Then

$$\begin{aligned} \varphi_X(t) &= \sum_{k=1}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} e^{itk} \\ &= e^{-\lambda} \sum_{k=1}^{\infty} \frac{(\lambda e^{it})^k}{k!} \\ &= e^{\lambda(e^{it} - 1)} \end{aligned}$$

(d) If X is uniformly distributed over $[a, b]$, then

$$\varphi_X(t) = \frac{e^{itb} - e^{ita}}{it(b-a)}$$

□

Characteristic functions and moments are related in the following way:

Proposition 10.2.7 *Suppose that a random variable X has an n^{th} moment, i.e. that*

$$\mathbb{E}|X|^n < +\infty$$

Then its characteristic function $\varphi_X(t)$ has a continuous derivative of order n , and

$$\varphi_X^{(n)}(t) = \mathbb{E}[(iX)^n e^{itX}]$$

Proof: Note that $\mathbb{E}[X^n e^{itX}]$ exists if and only if $\mathbb{E}|X|^n < +\infty$, i.e. if and only if X has an n^{th} moment.

We tackle first the case $n = 1$. Fix $t \in \mathbb{R}$, and suppose that $\mathbb{E}X$ exists. Then

$$\lim_{h \rightarrow 0} \frac{\varphi_X(t+h) - \varphi_X(t)}{h} = \lim_{h \rightarrow 0} \int e^{itx} \frac{e^{ihx} - 1}{h} \mu_X(dx)$$

But

$$\left| \int e^{itx} \frac{e^{ihx} - 1}{h} \right| \leq |t|$$

(You may wish to consult the “circle inequality” proved later in this chapter.) Thus we may apply the Lebesgue Dominated Convergence Theorem to deduce

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{\varphi_X(t+h) - \varphi_X(t)}{h} &= \int e^{itx} \lim_{h \rightarrow 0} \frac{e^{ihx} - 1}{h} \mu_X(dx) \\ &= \int ix e^{itx} \mu_X(dx) \end{aligned}$$

which proves

$$\varphi_X'(t) = i\mathbb{E}[X e^{itX}]$$

This yields the result for $n = 1$. Repeating the argument will establish the result for higher n .

◄

It follows that one may calculate the moments of a random variable directly from its characteristic function. Putting $t = 0$ in the above yields:

Proposition 10.2.8 *If X is a random variable with characteristic function $\varphi_X(t)$, then*

$$\mathbb{E}X^n = i^{-n} \varphi_X^{(n)}(0)$$

10.2.2 Inversion

The main result about characteristic functions is that random variables whose characteristic functions are equal are identically distributed, i.e. $\varphi_X = \varphi_Y$ if and only if $\mu_X = \mu_Y$. The distribution of a random variable can be recovered from its characteristic function. This result follows from the following theorem:

Theorem 10.2.9 (Lévy's Inversion Formula)

Let φ be the characteristic function of a random variable X , which has distribution μ and distribution function F . If $a < b \in \mathbb{R}$, then

$$\begin{aligned} \lim_{T \rightarrow +\infty} \frac{1}{2\pi} \int_{-T}^T \frac{e^{-ita} - e^{-itb}}{it} \varphi(t) dt \\ = \frac{1}{2} \mu(\{a\}) + \mu(a, b) + \frac{1}{2} \mu(\{b\}) \\ = \frac{1}{2} [F(b) + F(b-)] - \frac{1}{2} [F(a) + F(a-)] \end{aligned}$$

Remarks 10.2.10 Note that if F is continuous at a and b , then $\mu(\{a\}) = 0 = \mu(\{b\})$. In that case we have

$$\begin{aligned} \lim_{T \rightarrow +\infty} \frac{1}{2\pi} \int_{-T}^T \frac{e^{-ita} - e^{-itb}}{it} \varphi(t) dt \\ = \mu(a, b) \\ = F(b) - F(a) \end{aligned}$$

□

It follows from Lévy's Inversion Formula that the distribution of a random variable X is determined by its characteristic function. For suppose that X, Y are random variables which have the same characteristic function φ . Let μ_X be the distribution of X , and let μ_Y be the distribution of Y . Then for any real numbers $a < b$, we must have

$$\frac{1}{2} \mu_X(\{a\}) + \mu_X(a, b) + \frac{1}{2} \mu_X(\{b\}) = \frac{1}{2} \mu_Y(\{a\}) + \mu_Y(a, b) + \frac{1}{2} \mu_Y(\{b\})$$

There are at most countably many real places where a distribution function can be discontinuous. It therefore follows that there are sequences $a_n \downarrow a$ and $b_n \uparrow b$ such that both F_X, F_Y are continuous at each a_n and b_n . It follows by (a) that $\mu_X(a_n, b_n) = \mu_Y(a_n, b_n)$ for each n . Now since $(a_n, b_n) \uparrow (a, b)$, by continuity of measure we have

$$\mu_X(a, b) = \mu_Y(a, b)$$

Thus μ_X, μ_Y are finite measures which agree on all open intervals. But the open intervals form a π -system which generates the Borel algebra, and thus μ_X must agree with μ_Y on every Borel set. We have shown:

Theorem 10.2.11 Two random variables have the same characteristic functions if and only if they have the same distribution.

Examples 10.2.12 (a) If a random variable X has a characteristic function $\varphi_X(t) = e^{-\frac{t^2}{2}}$, then X must be normally distributed with mean 0 and variance 1. This follows from Example 7.3.8 and the Lévy Inversion Formula.

- (b) Suppose that X, Y are independent normally distributed random variables with mean 0 and variance 1. Then

$$\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t) = e^{-t^2}$$

Thus $X + Y$ has the same distribution function as a normally distributed random variable with mean 0 and variance 2. Hence $X + Y$ is a normally distributed random variable with mean 0 and variance 2.

- (c) If $\varphi_X(t)$ is *real-valued* (as opposed to complex-valued), then

$$\varphi_{-X}(t) = \overline{\varphi_X(t)} = \varphi_X(t)$$

Thus φ_X is real if and only if X and $-X$ are identically distributed, i.e. if and only if X is distributed symmetrically about the origin. As a consequence, all odd moments must be zero (if they exist).

□

Before we prove that Lévy's Inversion Formula holds, we need a definition and some lemmas:

Definition 10.2.13 Define the function $S(T)$ for $T \geq 0$ by

$$S(T) = \int_0^T \frac{\sin x}{x} dx$$

Also define the function $\text{sgn}(x)$ by

$$\text{sgn}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}$$

The function $x^{-1} \sin x$ itself is not Lebesgue integrable over the reals, because its positive and negative parts both integrate to infinity (exercise!). Nevertheless, $\lim_{T \rightarrow +\infty} \int_0^T x^{-1} \sin x dx$ exists, because the sequence

$$\left(\int_{(n-1)\pi}^{n\pi} x^{-1} \sin x dx \right)_n$$

is alternating and goes to 0 as $n \rightarrow +\infty$. It is important to know what this limit is:

Lemma 10.2.14

$$\lim_{T \rightarrow +\infty} \int_0^T \frac{\sin x}{x} dx = \frac{\pi}{2}$$

Proof: Note that

$$\int_0^T e^{-ux} \sin x dx = \frac{1}{1+u^2} [1 - e^{-uT}(u \sin T + \cos T)]$$

which may be obtained by two successive applications of integration by parts. Now the function $e^{-ux} \sin x$ is integrable over $(0, T) \times (0, +\infty)$, because the integral of its modulus is:

$$\int_0^T \int_0^\infty |e^{-ux} \sin x| du dx = \int_0^T x^{-1} |\sin x| dx \leq T < +\infty$$

using $|x^{-1} \sin x| \leq 1$. We may therefore apply Fubini's theorem:

$$\begin{aligned} \int_0^T \frac{\sin x}{x} dx &= \int_0^T \sin x \left[\int_0^\infty e^{-ux} du \right] dx \\ &= \int_0^\infty \left[\int_0^T e^{-ux} \sin x dx \right] du \\ &= \int_0^\infty \frac{du}{1+u^2} - \int_0^\infty \frac{e^{-uT}}{1+u^2} (u \sin T + \cos T) du \end{aligned}$$

Clearly $\int_0^\infty \frac{du}{1+u^2} = \frac{\pi}{2}$. Furthermore

$$\begin{aligned} 0 \leq \left| \int_0^\infty \frac{e^{-uT}}{1+u^2} (u \sin T + \cos T) du \right| &\leq \int_0^\infty \left| \frac{e^{-uT}}{1+u^2} \right| |u \sin T + \cos T| du \\ &\leq \int_0^\infty e^{-uT} (u+1) du \rightarrow 0 \quad \text{as } T \rightarrow \infty \end{aligned}$$

by the Lebesgue Dominated Convergence Theorem, so that $\int_0^\infty \frac{e^{-uT}}{1+u^2} (u \sin T + \cos T) du \rightarrow 0$ as $T \rightarrow +\infty$. This is what we needed to establish.

—

Another technical result that we shall need is the following:

Lemma 10.2.15

$$\frac{1}{2\pi} \int_{-T}^T \frac{e^{itx}}{it} dt = \frac{\operatorname{sgn}(x) S(|x|T)}{\pi}$$

Proof:

$$\int_{-T}^T \frac{\cos tx + i \sin tx}{it} dt = \int_{-T}^T \frac{\sin tx}{t} dt$$

because $\frac{\cos tx}{it}$ is an odd function of t . Moreover, using the fact that $\frac{\sin tx}{t}$ is even and a change of variables $t \mapsto \frac{u}{x}$, we obtain

$$\int_{-T}^T \frac{\sin tx}{t} dt = 2 \operatorname{sgn}(x) \int_0^{T|x|} \frac{\sin u}{u} du$$

from which the result now follows trivially.

—

The final lemma will be useful in giving upper bounds for certain functions:

Lemma 10.2.16 (“circle inequality”)

For $u, v \in \mathbb{R}$ with $u \leq v$ we have

$$|e^{iv} - e^{iu}| \leq |v - u|$$

Proof: This follows from the fact that

$$\left| \int_u^v i e^{it} dt \right| \leq \int_u^v |i e^{it}| dt = v - u$$

(or more simply from a diagram: Just draw e^{iv} and e^{iu} on the unit circle in the complex plane and think about it.)

+

We may now prove the Lévy Inversion Formula:

Proof of Lévy's Inversion Formula:

If $a < b$ in \mathbb{R} , and if $0 \leq T < +\infty$, then

$$\begin{aligned} & \frac{1}{2\pi} \int_{-T}^T \frac{e^{-ita} - e^{-itb}}{it} \varphi(t) dt \\ &= \frac{1}{2\pi} \int_{-T}^T \frac{e^{-ita} - e^{-itb}}{it} \left(\int_{\mathbb{R}} e^{itx} \mu(dx) \right) dt \\ &= \frac{1}{2\pi} \int_{\mathbb{R}} \left(\int_{-T}^T \frac{e^{it(x-a)} - e^{it(x-b)}}{it} dt \right) \mu(dx) \end{aligned}$$

by Fubini's Theorem, since the integral of the absolute value is finite: By Tonelli's Theorem,

$$\begin{aligned} & \frac{1}{2\pi} \int_{[-T, T] \times \mathbb{R}} \left| \frac{e^{it(x-a)} - e^{it(x-b)}}{it} \right| (\lambda \otimes \mu)(dt, dx) \\ &= \frac{1}{2\pi} \int_{\mathbb{R}} \left(\int_{-T}^T \left| \frac{e^{it(x-a)} - e^{it(x-b)}}{it} \right| dt \right) \mu(dx) \\ &\leq \frac{1}{2\pi} \int_{\mathbb{R}} \left(\int_{-T}^T |b-a| dt \right) \mu(dx) \\ &= \frac{(b-a)T}{\pi} \end{aligned}$$

using the inequality on the unit circle above.

Using one of the lemmas above, we obtain

$$\begin{aligned} & \frac{1}{2\pi} \int_{-T}^T \frac{e^{it(x-a)} - e^{it(x-b)}}{it} dt \\ &= \frac{\operatorname{sgn}(x-a)S(|x-a|T) - \operatorname{sgn}(x-b)S(|x-b|T)}{\pi} \end{aligned}$$

Now as $T \uparrow +\infty$, $S(T) \rightarrow \frac{\pi}{2}$, and thus

$$\lim_{T \rightarrow +\infty} \frac{\operatorname{sgn}(x-a)S(|x-a|T) - \operatorname{sgn}(x-b)S(|x-b|T)}{\pi} = \begin{cases} 0 & \text{if } x < a \\ \frac{1}{2} & \text{if } x = a \\ 1 & \text{if } a < x < b \\ \frac{1}{2} & \text{if } x = b \\ 0 & \text{if } b < x \end{cases}$$

Hence

$$\lim_{T \rightarrow +\infty} \frac{\operatorname{sgn}(x-a)S(|x-a|T) - \operatorname{sgn}(x-b)S(|x-b|T)}{\pi} = \frac{1}{2}I_{\{a\}} + I_{(a,b)} + \frac{1}{2}I_{\{b\}}$$

It follows that

$$\begin{aligned} \lim_{T \rightarrow +\infty} \frac{1}{2\pi} \int_{-T}^T \frac{e^{-ita} - e^{-itb}}{it} \varphi(t) dt \\ &= \lim_{T \rightarrow +\infty} \int_{\mathbb{R}} \frac{\operatorname{sgn}(x-a)S(|x-a|T) - \operatorname{sgn}(x-b)S(|x-b|T)}{\pi} \mu(dx) \\ &= \int_{\mathbb{R}} \lim_{T \rightarrow +\infty} \frac{\operatorname{sgn}(x-a)S(|x-a|T) - \operatorname{sgn}(x-b)S(|x-b|T)}{\pi} \mu(dx) \\ &= \frac{1}{2}\mu(\{a\}) + \mu(a,b) + \frac{1}{2}\mu(\{b\}) \end{aligned}$$

where we used the Lebesgue Dominated Convergence Theorem to take the limit inside the integral. This proves the result.

Note also that

$$\frac{1}{2}[F(b) + F(b-) - F(a) - F(a-)] = \frac{1}{2}[\mu(a, b] + \mu[a, b]]$$

□

An extremely useful (but slightly weaker) version of the inversion theorem is the following:

Theorem 10.2.17 *Let φ be the characteristic function of a random variable X , which has distribution μ and distribution function F . If*

$$\int_{\mathbb{R}} |\varphi(t)| dt < +\infty$$

then X has a continuous probability density function f_X , and

$$f_X(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-itx} \varphi(t) dt$$

Exercise 10.2.18 We prove Thm. 10.2.17

(a) Show that $\int_{\mathbb{R}} \left| \frac{e^{-ita} - e^{-itb}}{it} \varphi(t) \right| dt \leq |b-a| \int_{\mathbb{R}} |\varphi(t)| dt$ and conclude that $\frac{e^{-ita} - e^{-itb}}{it} \varphi(t)$ is integrable over \mathbb{R} (with respect to t).

(b) Explain why

$$F(b) - F(a) \leq \frac{|b-a|}{2\pi} \int_{\mathbb{R}} |\varphi(t)| dt$$

and why this, together with the fact that $\int_{\mathbb{R}} |\varphi(t)| dt < +\infty$, immediately implies that F is continuous.

(c) Explain why

$$F(b) - F(a) = \frac{1}{2\pi} \int_{\mathbb{R}} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t) dt$$

for all $a < b \in \mathbb{R}$.

(d) Deduce that that for all x and all h , we have

$$\frac{F(x+h) - F(x)}{h} = \frac{1}{2\pi} \int_{\mathbb{R}} \frac{e^{-itx} - e^{-it(x+h)}}{ith} \varphi(t) dt$$

(e) Use the Lebesgue Dominated Convergence Theorem to conclude that

$$f_X(x) = F'(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-itx} \varphi(t) dt$$

as required.

(f) Finally, explain why f_X must be a continuous function.

□

Remarks 10.2.19 If X is a random with a continuous probability density function f_X , then we have the following “duality”:

$$\begin{aligned} \varphi_X(t) &= \int_{\mathbb{R}} f_X(x) e^{itx} dx \\ f_X(x) &= \frac{1}{2\pi} \int_{\mathbb{R}} \varphi_X(t) e^{-itx} dt \end{aligned}$$

Thus the density and characteristic functions are inverses of each other under Fourier transforms.

□

10.2.3 Weak Convergence and Characteristic Functions

Our next result proves that weak convergence of measures is equivalent to pointwise convergence of characteristic functions:

Theorem 10.2.20 (Lévy Continuity Theorem)

Let μ_n be probability measures on $(\mathbb{R}, \mathcal{B})$, and let φ_n be the associated characteristic functions. Suppose that $(\varphi_n(t))_n$ converges for every $t \in \mathbb{R}$, and that

$$\psi(t) := \lim_n \varphi_n(t)$$

If φ is continuous at $t = 0$, then ψ is the characteristic function of a probability distribution μ , and $\mu_n \xrightarrow{w} \mu$

Proof: Suppose first that $(\mu_n)_n$ is tight. Then by the Helly–Bray Lemma there is a subsequence $(\mu_{n_k})_k$ and a probability distribution μ such that $\mu_{n_k} \xrightarrow{w} \mu$. Now $e^{itx} = \cos tx + i \sin tx$ is continuous, and hence, by definition of weak convergence, we have $\int e^{itx} \mu_{n_k}(dx) \rightarrow \int e^{itx} \mu(dx)$, i.e. $\varphi_{n_k}(t) \rightarrow \varphi(t)$, where φ is the characteristic function of μ . But as $\varphi_n(t) \rightarrow \psi(t)$, the same is true for any subsequence. Hence $\varphi(t) = \psi(t)$.

So far, we know only that $\mu_{n_k} \xrightarrow{w} \mu$, but we would like to show that $\mu_n \xrightarrow{w} \mu$. It is easier to argue with distribution functions: If $\mu_n \not\xrightarrow{w} \mu$ does not converge weakly to μ , then $F_n \not\xrightarrow{d} F$ (where, of course, F_n, F are the distribution functions of μ_n, μ), which means that there is a continuity point x of F such that $F_n(x) \not\rightarrow F(x)$. It follows that there is a $\varepsilon > 0$ and a subsequence $(F_{m_j})_j$ such that $|F_{m_j}(x) - F(x)| \geq \varepsilon$ for all j . Applying Helly–Bray Lemma to this subsequence, we obtain a subsubsequence $(F_{m_{j_l}})_l$ which converges weakly, by tightness, to some distribution function G , with characteristic function φ_G . But since $\varphi_n(t) \rightarrow \psi(t)$, we must have $\varphi_{m_{j_l}}(t) \rightarrow \psi(t)$, and hence $\varphi_G(t) = \psi(t) = \varphi(t)$. Thus G and F have the same

characteristic function, and hence $G = F$, by Lévy's Inversion Theorem. But this leads to a contradiction: As $F_{m_j}(x) \rightarrow G(x)$ and $G(x) = F(x)$, we have $F_{m_{j_l}}(x) \rightarrow F(x)$ as $l \rightarrow \infty$. Yet $|F_{m_j}(x) - F(x)| \geq \varepsilon$ for all $j \in \mathbb{N}$, and thus we have both

$$|F_{m_{j_l}}(x) - F(x)| \geq \varepsilon \text{ for all } l \quad \text{and} \quad |F_{m_{j_l}}(x) - G(x)| < \varepsilon \text{ for all large } l$$

which is impossible.

It just remains to show that the $(\mu_n)_n$ are tight: Now if $\delta > 0$, then

$$\begin{aligned} \delta^{-1} \int_{-\delta}^{\delta} 1 - \varphi_n(t) dt &= \delta^{-1} \int_{-\delta}^{\delta} \int_{\mathbb{R}} 1 - e^{itx} \mu_n(dx) dt \\ &= \int_{\mathbb{R}} \delta^{-1} \int_{-\delta}^{\delta} 1 - e^{itx} dt \mu_n(dx) \\ &= 2 \int_{\mathbb{R}} \left(1 - \frac{\sin \delta x}{\delta x}\right) \mu_n(dx) \\ &\geq 2 \int_{|x| > 2\delta^{-1}} \left(1 - \frac{1}{|\delta x|}\right) \mu_n(dx) \\ &\geq \mu_n(\{x : |x| > 2\delta^{-1}\}) \end{aligned}$$

Here we used Fubini's Theorem to change the order of integration (which is permitted, since $1 - e^{itx}$ is bounded by 2). We also used the fact that $1 - \frac{\sin \delta x}{\delta x} \geq 1 - \frac{1}{|\delta x|}$ and that $1 - \frac{1}{|\delta x|} > \frac{1}{2}$ when $|x| > 2\delta^{-1}$.

Now clearly $\psi(0) = \lim \varphi_n(0) = \lim_n 1 = 1$. As ψ is assumed to be continuous at $t = 0$, there is for every $\varepsilon > 0$ a $\delta > 0$ such that

$$\delta^{-1} \int_{-\delta}^{\delta} 1 - \varphi(t) dt < \varepsilon$$

Since $\varphi_n(t) \rightarrow \varphi(t)$, the Lebesgue Dominated Convergence Theorem implies that there exists an N such that

$$\delta^{-1} \int_{-\delta}^{\delta} 1 - \varphi_n(t) dt < \varepsilon \quad \text{for all } n \geq N$$

It follows that

$$\mu_n[-2\delta^{-1}, 2\delta^{-1}] > 1 - \varepsilon$$

for all $n \geq N$. Also pick $K_1, K_2, \dots, K_N > 0$ such that

$$\mu_n[-K_n, K_n] > 1 - \varepsilon \quad \text{for } n = 1, \dots, N$$

If $K := \max\{2\delta^{-1}, K_1, K_2, \dots, K_N\}$, then

$$\mu_n[-K, K] > 1 - \varepsilon \quad \text{for all } n \in \mathbb{N}$$

proving that $(\mu_n)_n$ is indeed tight.

Now if μ_n, μ are probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, with characteristic functions φ_n, φ , and $\mu_n \xrightarrow{w} \mu$, then we must certainly have $\int e^{itx} \mu_n(dx) \rightarrow \int e^{itx} \mu(dx)$, by definition of weak convergence, as $e^{itx} = \cos tx + i \sin tx$ is continuous. Thus

$$\mu_n \xrightarrow{w} \mu \quad \text{implies} \quad \varphi_n(t) \rightarrow \varphi(t) \text{ for all } t \in \mathbb{R}$$

Conversely, if $\varphi_n(t) \rightarrow \varphi(t)$ for all $t \in \mathbb{R}$, then, by the previous theorem, $\mu_n \xrightarrow{w} \mu$, as φ — being a characteristic function — is continuous, and thus continuous at $t = 0$. We thus have:

Corollary 10.2.21 *Suppose that μ_n, μ are probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, with characteristic functions φ_n, φ . Then*

$$\mu_n \xrightarrow{w} \mu \quad \text{if and only if} \quad \varphi_n(t) \rightarrow \varphi(t) \text{ for all } t \in \mathbb{R}$$

In the proof of the Lévy Continuity Theorem, we proved the following useful inequality:

Proposition 10.2.22 *If φ is the characteristic function of a probability measure μ on \mathbb{R} , and if $\delta > 0$, then*

$$\delta^{-1} \int_{-\delta}^{\delta} 1 - \varphi(t) dt \geq \mu(\{x : |x| \geq 2\delta^{-1}\})$$

□

10.3 The Central Limit Theorem

The Central Limit Theorem is one of the fundamental results in mathematics, and is easy to prove with the machinery set up in the previous two sections. It states that if X_n is a sequence of independent identically distributed random variables with mean μ and variance σ^2 , then the distribution of the fractions $(X_1 \cdots + X_n - n\mu)/\sigma\sqrt{n}$ tends to a standard normal distribution (with mean 0 and variance 1). Why the \sqrt{n} ? Basically, if we define $S_n = X_1 + \cdots + X_n$, then it is clear that S_n has mean $n\mu$ and variance $n\sigma^2$. Thus as $n \rightarrow +\infty$, $\text{var}(S_n) \rightarrow +\infty$ as well. However, S_n/\sqrt{n} has variance $\text{var}(S_n)/n = \sigma^2$.

We can conclude that if n is sufficiently large, then S_n is approximately normally distributed with mean $n\mu$ and variance $n\sigma^2$. Thus the sums of identically distributed random variables tend to become normally distributed.

Here is a concrete exercise which contains all the important elements of the proof of the Central Limit Theorem:

Exercise 10.3.1 (a) Suppose that X is a random variable with

$$\mathbb{P}(X = 1) = \frac{1}{2} = \mathbb{P}(X = -1)$$

and that φ_X is characteristic function of X . Show that

$$\varphi_X(u) = \cos u$$

(b) Now let X_n (for $n \in \mathbb{N}$) be independent random variables, all with the same distribution as X in (a). For $n \in \mathbb{N}$, define random variables G_n by

$$G_n := \frac{X_1 + X_2 + \cdots + X_n}{\sqrt{n}}$$

Show that the characteristic function φ_{G_n} of G_n is given by

$$\varphi_{G_n}(u) = \left(\cos \frac{u}{\sqrt{n}}\right)^n$$

(c) Use a Taylor expansion of $\cos x$ about $x = 0$ to show that

$$\cos \frac{u}{\sqrt{n}} = 1 + \frac{u^2}{n} \left(-\frac{1}{2} + \varepsilon\left(\frac{u}{\sqrt{n}}\right)\right) \quad \text{where } \varepsilon(h) \rightarrow 0 \text{ as } h \rightarrow 0$$

(d) Now define $k := k\left(\frac{u}{\sqrt{n}}\right) := -\frac{1}{2} + \varepsilon\left(\frac{u}{\sqrt{n}}\right)$ so that

$$k \rightarrow -\frac{1}{2} \quad \text{as } n \rightarrow \infty \quad \text{and} \quad \cos \frac{u}{\sqrt{n}} = 1 + \frac{u^2 k}{n}$$

By considering $\ln \varphi_{G_n}(u)$, and with the aid of a Taylor expansion of $\ln(1+x)$ about $x = 0$, show that

$$\lim_{n \rightarrow \infty} \varphi_{G_n}(u) = e^{-\frac{1}{2}u^2}$$

(e) Now explain why we may now deduce that $G_n \xrightarrow{w} Z$, where Z is a standard normal random variable.

□

Theorem 10.3.2 (Central Limit Theorem)

Let N be a standard normally distributed random variable, with mean 0 and variance 1. Let X_n be a sequence of independent identically distributed random variables with mean μ and variance $\sigma^2 < +\infty$. Define $S_n = X_1 + X_2 + \cdots + X_n$ and set

$$G_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

Then

$$G_n \xrightarrow{w} N$$

Proof: It clearly suffices to prove the theorem for independent identically distributed X_n with mean zero and variance 1, because if the X_n have mean μ and variance σ^2 , then $\tilde{X}_n := (X_n - \mu)/\sigma$ have mean zero and variance 1, and then

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\tilde{S}_n}{\sqrt{n}}$$

So let $(X_n)_n$ be a sequence of independent random variables with mean zero and variance 1. Define

$$G_n = \frac{X_1 + X_2 + \cdots + X_n}{\sqrt{n}}$$

Then because the X_k are independent, we can say that

$$\varphi_{G_n}(t) = \prod_{k=1}^n \varphi_{\frac{X_k}{\sqrt{n}}}(t) = \varphi\left(\frac{t}{\sqrt{n}}\right)^n$$

where φ is the common characteristic function of the X_k . Now if we consider a Taylor expansion for φ about $t = 0$, we'll obtain something like

$$\varphi(t) = \varphi(0) + \varphi'(0)t + \frac{1}{2}\varphi''(0)t^2 + \varepsilon(t)t^2$$

where $\varepsilon(t) \rightarrow 0$ as $t \rightarrow 0$. Taylor's Theorem applies because X has a second moment, so that φ is twice differentiable. Now examining the moments, we obtain

$$\varphi(0) = 1 \quad \varphi'(0) = 0 \quad \varphi''(0) = -1$$

Putting this back into the Taylor expansion for φ yields

$$\varphi(t) = 1 - \frac{1}{2}t^2 + \varepsilon(t)t^2 = 1 + [-\frac{1}{2} + \varepsilon(t)]t^2 = 1 + k(t)t^2$$

where $k(t) \rightarrow -\frac{1}{2}$ as $t \rightarrow 0$.

Thus

$$\varphi_{G_n}(t) = \varphi\left(\frac{t}{\sqrt{n}}\right)^n = \left[1 + \frac{k(\frac{t}{\sqrt{n}})t^2}{n}\right]^n$$

Now recall that the first order Taylor expansion of $\ln(1+z)$ about $x = 0$ yields

$$\ln(1+z) = z + \bar{\varepsilon}(z)|z| \quad \text{where } \bar{\varepsilon}(z) \rightarrow 0 \text{ as } z \rightarrow 0$$

Put $z := \frac{k(\frac{t}{\sqrt{n}})t^2}{n}$, so that $z \rightarrow 0$ as $n \rightarrow \infty$, to obtain

$$\begin{aligned} \ln \varphi_{G_n}(t) &= n \ln \left(1 + \frac{k(\frac{t}{\sqrt{n}})t^2}{n}\right) \\ &= n \ln(1+z) \\ &= n(z + \bar{\varepsilon}(z)|z|) \\ &= k(\frac{t}{\sqrt{n}})t^2 + \bar{\varepsilon}(z)k(\frac{t}{\sqrt{n}})t^2 \end{aligned}$$

and hence

$$\lim_n \ln \varphi_{G_n}(t) = \lim_n k(\frac{t}{\sqrt{n}})t^2 + \lim_n \bar{\varepsilon}(z)k(\frac{t}{\sqrt{n}})t^2 = -\frac{1}{2}t^2$$

It therefore follows that

$$\varphi_{G_n}(t) \rightarrow e^{-t^2/2} \text{ for all } t$$

By Lévy's Convergence Theorem, the distributions μ_n of G_n converge weakly to a distribution whose characteristic function is $e^{-t^2/2}$. But this is the characteristic function of the standard normal distribution, so by Lévy's Inversion Formula it follows that G_n weakly converges to a standard normally distributed random variable.

—

Chapter 11

Conditional Expectation and Martingales

11.1 Information and Expectation

Because a thorough understanding of conditioning is absolutely fundamental to the development and understanding of martingales, stochastic integrals and the tools of arbitrage pricing, we take another, gentler, look at how information is organised and used in probability theory. It is clear that new information will cause a re-evaluation of probabilities, and thus the expected values of random variables. Information does not arrive all at once, but in dribs and drabs over time. Information is contained in σ -algebras, and the flow of information will be modelled by an increasing chain of σ -algebras. Given a random variable such as the stock price S_T at some later time T , we have an initial expectation $\mathbb{E}_0 S_T$ of what the stock price will be, but as we observe the stock price over the interval $[0, T]$, our expectation changes: We get a sequence $\mathbb{E}_t S_T$ of expectations at time t . At time T , we know the stock price exactly, and so $\mathbb{E}_T S_T = S_T$, i.e. the expected value at time T is the actual value. It will turn out that under an equivalent martingale measure we have

$$\mathbb{E}_t \bar{S}_T = \bar{S}_t$$

i.e. at time t , the (risk-neutral) expected value of discounted S_T is just the discounted value of S_t . Here the discounting is done back to the period $t = 0$.

The right tool to deal with these problems is the notion of *conditional expectation*, due to Kolmogorov. This is regarded by many as *the* central concept of probability theory.

11.1.1 Conditioning on an Event

The first, and simplest, case that we consider is that of the conditional expectation $\mathbb{E}(X|A)$ of a random variable given an event A . Recall that $\mathbb{E}X$ is just the Lebesgue integral of X with respect to the probability measure \mathbb{P} , and that this essentially amounts to taking a weighted sum of the values of the random variables, where the weights are the probabilities: In the discrete framework, with $\Omega = \{\omega_n : n = 1, 2, \dots\}$, we have

$$\mathbb{E}(X) = \int_{\Omega} X \, d\mathbb{P} = \sum_n \mathbb{P}(\{\omega_n\}) X(\omega_n)$$

This makes sense from the frequentists' point of view: If we perform the random experiment a large number N of times, the outcome ω_n occurs roughly $N\mathbb{P}(\{\omega_n\})$ —many times. Thus we observe a value of $X(\omega_n)$ roughly $N\mathbb{P}(\{\omega_n\})$ times. The average value of X is just the sum over all experiments of all the values of X divided by the number of experiments, i.e.

$$\frac{1}{N} \sum_n N\mathbb{P}(\{\omega_n\})X(\omega_n) = \mathbb{E}X$$

Note that we can also write this as

$$\mathbb{E}X = \sum_{\omega \in \Omega} \mathbb{P}(\{\omega\})X(\omega)$$

It is just the weighted average of X over the set Ω . It will often be useful to take the weighted average of X over a subset B of Ω . To that end we define:

$$\mathbb{E}(X; B) = \int_B X \, d\mathbb{P} = \mathbb{E}[XI_B]$$

(Recall that $\int_B X \, d\mathbb{P}$ is *defined* as $\int XI_B \, d\mathbb{P}$. Since I_B equals 1 on B , and 0 outside B , this corresponds to finding the weighted average of X , but only for those values for which the outcome lies in B .)

In the discrete framework we obviously have:

$$\mathbb{E}(X; B) = \sum_{\omega \in B} \mathbb{P}(\{\omega\})X(\omega)$$

Note also that $\mathbb{E}X = \mathbb{E}(X; \Omega)$.

Suppose now that we are given the information that the event A occurred. In that case the only possible values of X that we will observe are of the form $X(\omega)$, $\omega \in A$. We will therefore not generally be able to observe all the possible values of X , but just those $X(\omega)$ for which $\omega \in A$.

How should we define $\mathbb{E}(X|A)$, the expected value of X given A ? Let's try the frequentists' approach for guidance: Suppose that a superlarge number M of random trials is performed. The event A won't occur every single time, but if M is superlarge, there should still be a large number $N \approx M\mathbb{P}(A)$ of times in which the outcome does lie in A .¹ If $\omega \in A$, we will therefore see an outcome $X(\omega)$ roughly $M\mathbb{P}(\{\omega\})$ —many times. Thus the average value of X given that A occurs is simply the sum over all experiments for which A occurs, divided by the number of times that A occurs. Thus we ought to have

$$\begin{aligned} \mathbb{E}(X|A) &= \frac{1}{N} \sum_{\omega \in A} M\mathbb{P}(\{\omega\})X(\omega) \\ &= \sum_{\omega \in A} \frac{\mathbb{P}(\{\omega\})X(\omega)}{\mathbb{P}(A)} \\ &= \frac{\mathbb{E}(X; A)}{\mathbb{P}(A)} \end{aligned}$$

We therefore define:

¹Unless $\mathbb{P}(A) = 0$.

Definition 11.1.1 If X is a random variable and if A is an event with $\mathbb{P}(A) \neq 0$, then the **conditional expectation of X given A** is defined by:

$$\mathbb{E}(X|A) = \frac{\mathbb{E}(X; A)}{\mathbb{P}(A)}$$

□

Example 11.1.2 Suppose that X is a random variable and that A is an event with positive probability. Before we know that A has occurred, we have a best estimate of what x will be, namely $\mathbb{E}X$. If we are told that A has occurred, however, we will revise our expectation. For example, the expected value if a fair die is rolled is 3.5. However, if we are told that the outcome is even, we revise our estimate: the expected value is now 4. If we are told that the outcome is odd, the expected value is 3. Essentially, when we are told that an event A has occurred, we revise our probability measure: Each outcome outside A is now assigned measure 0, whereas each event inside A has its probability scaled up by dividing by $\mathbb{P}(A)$: $\mathbb{P}(B|A) = \frac{\mathbb{P}(B)}{\mathbb{P}(A)}$ for all $B \subseteq A$. We use the new measure $\mathbb{P}(\cdot|A)$ on the sample space A to calculate our revised expectation: The expectation of X given A is

$$\begin{aligned} \mathbb{E}[X|A] &= \int_A X \, d\mathbb{P}(\cdot|A) = \frac{1}{\mathbb{P}(A)} \int_A X \, d\mathbb{P} \\ &= \frac{\mathbb{E}[X; A]}{\mathbb{P}(A)} \end{aligned}$$

Note that $\mathbb{P}(B) = \mathbb{E}I_B$ for every event B . Generalizing, we can define the **conditional probability** of the event B given A by

$$\mathbb{P}(B|A) = \mathbb{E}(I_B|A)$$

□

11.1.2 Conditioning on a Random Variable

Next we tackle a slightly more complicated version of conditional expectation, namely the conditional expectation of the random variable X given the random variable Y , denoted $\mathbb{E}(X|Y)$. The idea is that if we know the values of the random variable Y , this may give us information about the random variable X , unless they are independent.

Suppose first that we are working with a discrete probability space $\Omega = \{\omega_n : n = 1, 2, \dots\}$. Let $\{y_m : m = 1, 2, \dots\}$ be the set of values of Y , where we assume that $\mathbb{P}(Y = y_n) \neq 0$. We thus have a sequence of events $\{Y = y_m\}$, and for each of these we can calculate the conditional expectation $\mathbb{E}(X|Y = y_m)$. Instead of regarding each of the

$$\mathbb{E}(X|Y = y_1), \mathbb{E}(X|Y = y_2), \mathbb{E}(X|Y = y_3), \dots$$

separately, we define instead a new random variable $\mathbb{E}(X|Y)$ as follows:

$$\begin{aligned} \mathbb{E}(X|Y)(\omega) &= \mathbb{E}(X|Y = Y(\omega)) \\ \text{i.e.} \quad \mathbb{E}(X|Y)(\omega) &= \mathbb{E}(X|Y = y_m) \quad \text{if } Y(\omega) = y_m \end{aligned}$$

Example 11.1.3 A fair die is rolled, and you get an amount equal to the outcome ω , but only if it is even. Let X be your winnings. Let Y, Z be defined as follows:

$$Y(\omega) = \begin{cases} 1 & \text{if } \omega \text{ is even} \\ 0 & \text{else} \end{cases} \quad Z(\omega) = \begin{cases} 0.1234 & \text{if } \omega \text{ is even} \\ 10\sqrt{\pi} & \text{else} \end{cases}$$

We calculate the random variable $\mathbb{E}[X|Y]$:

$$\begin{aligned} \mathbb{E}[X|Y](\omega) &= \begin{cases} \frac{\mathbb{E}[X; Y = 1]}{\mathbb{P}(Y = 1)} & \text{if } Y(\omega) = 1 \\ \frac{\mathbb{E}[X; Y = 0]}{\mathbb{P}(Y = 0)} & \text{if } Y(\omega) = 0 \end{cases} \\ &= \begin{cases} \frac{\mathbb{E}[X; \{2, 4, 6\}]}{\mathbb{P}(\{2, 4, 6\})} & \text{if } \omega \in \{2, 4, 6\} \\ \frac{\mathbb{E}[X; \{1, 3, 5\}]}{\mathbb{P}(\{2, 4, 6\})} & \text{if } \omega \in \{2, 4, 6\} \end{cases} \\ &= \begin{cases} 4 & \text{if } \omega \text{ is even} \\ 0 & \text{if } \omega \text{ is odd} \end{cases} \end{aligned}$$

Now compare $\mathbb{E}[X|Y]$ and $\mathbb{E}[X|Z]$:

$$\begin{aligned} \mathbb{E}[X|Y](\omega) &= \begin{cases} \frac{\mathbb{E}[X; Y = 1]}{\mathbb{P}(Y = 1)} & \text{if } Y(\omega) = 1 \\ \frac{\mathbb{E}[X; Y = 0]}{\mathbb{P}(Y = 0)} & \text{if } Y(\omega) = 0 \end{cases} \\ &= \begin{cases} \frac{\mathbb{E}[X; \{2, 4, 6\}]}{\mathbb{P}(\{2, 4, 6\})} & \text{if } \omega \in \{2, 4, 6\} \\ \frac{\mathbb{E}[X; \{1, 3, 5\}]}{\mathbb{P}(\{2, 4, 6\})} & \text{if } \omega \in \{2, 4, 6\} \end{cases} \\ &= \begin{cases} \frac{\mathbb{E}[X; Z = 0.1234]}{\mathbb{P}(Z = 0.1234)} & \text{if } Z(\omega) = 0.1234 \\ \frac{\mathbb{E}[X; Z = 10\sqrt{\pi}]}{\mathbb{P}(Z = 10\sqrt{\pi})} & \text{if } Z(\omega) = 10\sqrt{\pi} \end{cases} \\ &= \mathbb{E}[X|Z](\omega) \end{aligned}$$

Even though Y, Z are quite different, we see that $\mathbb{E}[X|Y] = \mathbb{E}[X|Z]$. This is because the conditional expectation $\mathbb{E}[X|Y]$ depends not on the *values* of Y , but on the *information* in Y , i.e. it depends on $\sigma(Y)$. It is obvious, in this case, that

$$\sigma(Y) = \sigma(\{1, 3, 5\}, \{2, 4, 6\}) = \sigma(Z)$$

i.e. Y, Z contain the same information.

□

Now note that $\sigma(Y)$ is obviously generated by the sets partition $\{Y = y_1\}, \{Y = y_2\}, \dots$, and that $\mathbb{E}(X|Y)$ is constant on each block (i.e. that $\mathbb{E}(X|Y)(\omega_1) = \mathbb{E}(X|Y)(\omega_2)$ if ω_1, ω_2 belong to the same block; this happens if $Y(\omega_1) = Y(\omega_2)$). It follows that

$$\mathbb{E}[X|Y] \text{ is } \sigma(Y)\text{-measurable}$$

The intuition behind this fact is simple: If we know Y , we can calculate $\mathbb{E}(X|Y)$. Thus all the information needed to calculate $\mathbb{E}(X|Y)$ is contained in $\sigma(Y)$. We must therefore have $\sigma(\mathbb{E}(X|Y)) \subseteq \sigma(Y)$.

Now each element of $\sigma(Y)$ is simply a finite union of the sets $\{Y = y_m\}$ which make up the partition which generates $\sigma(Y)$. Since $\mathbb{E}[X|Y]$ is a random variable, we can integrate it. Note that on each block we have:

$$\begin{aligned} \mathbb{E}[\mathbb{E}[X|Y]; Y = y_m] &= \mathbb{E}[X|Y = y_m] \cdot \mathbb{P}(Y = y_m) \\ &= \frac{\mathbb{E}[X; Y = y_m]}{\mathbb{P}[Y = y_m]} \mathbb{P}[Y = y_m] \\ &= \mathbb{E}[X; Y = y_m] \end{aligned}$$

[Since $\mathbb{E}(X; A) = \int_A X \, d\mathbb{P}$ is the “average” of the random variable X over the set A , it follows that $X, \mathbb{E}[X|Y]$ have the same average values over each set $\{Y = y_m\}$. In different notation:

$$\int_{\{Y=y_m\}} \mathbb{E}[X|Y] \, d\mathbb{P} = \int_{\{Y=y_m\}} X \, d\mathbb{P}$$

Again, this simply states that the random variables X and $\mathbb{E}[X|Y]$ have the same average over any set in the partition that generates $\sigma(Y)$. Since $\int_{F_1 \cup F_2} X \, d\mathbb{P} = \int_{F_1} X \, d\mathbb{P} + \int_{F_2} X \, d\mathbb{P}$ if F_1, F_2 are disjoint, it now follows readily that $X, \mathbb{E}[X|Y]$ have the same average over any set in $\sigma(Y)$: The elements of $\sigma(Y)$ are just disjoint unions of the blocks $\{Y = y_m\}$.

Thus we have shown the following:

Proposition 11.1.4 Let X be a *discrete* integrable random variable^a and let Y be an arbitrary random variable. Then **conditional expectation** $\mathbb{E}[X|Y]$ is a random variable with the properties that

- (a) $\mathbb{E}[X|Y]$ is $\sigma(Y)$ -measurable.
- (b) $\int_A \mathbb{E}[X|Y] \, d\mathbb{P} = \int_A X \, d\mathbb{P}$ for all $A \in \sigma(Y)$.

□

^ai.e. $\mathbb{E}X$ exists.

The above proposition makes it clear that $\mathbb{E}[X|Y]$ depends on Y only through $\sigma(Y)$, and not directly on the values of Y , i.e. that

Corollary 11.1.5 Suppose that X, Y, Z are random variables, with Y, Z discrete. If $\sigma(Y) = \sigma(Z)$ then $\mathbb{E}[X|Y] = \mathbb{E}[X|Z]$ a.s.

□

We have not, as yet, proved that $\mathbb{E}[X|Y]$ exists for general Y , but only for discrete Y (i.e. Y with at most countably many values). However, as the definition of $\mathbb{E}[X|Y]$ depends on Y only through $\sigma(Y)$, we proceed to generalize:

11.1.3 Conditioning on a σ -Algebra

Definition and Theorem 11.1.6 (Kolmogorov)

Suppose that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and that X is a random variable in $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$. Let \mathcal{G} be a sub- σ -algebra of \mathcal{F} . Then there exists a random variable Z such that

- (i) Z is \mathcal{G} -measurable.
- (ii) $Z \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, i.e. $\mathbb{E}(|Z|) < +\infty$.
- (iii) For every set $G \in \mathcal{G}$, we have

$$\mathbb{E}[Z; G] = \mathbb{E}[X; G] \quad \text{i.e.} \quad \int_G Z \, d\mathbb{P} = \int_G X \, d\mathbb{P}$$

Moreover, if Z' is a random variable satisfying (i),(ii),(iii), then $Z = Z'$ a.s. Any random variable Z with the properties (i),(ii),(iii) is called a *version* of the conditional expectation of X given \mathcal{G} . We write

$$Z = \mathbb{E}[X|\mathcal{G}] \quad \text{a.s.} \quad \text{or} \quad Z = \mathbb{E}^{\mathcal{G}} X \quad \text{a.s.}$$

Definition 11.1.7 We define conditional expectation w.r.t to general random variables in the following manner:

$$\mathbb{E}[X|Y] := \mathbb{E}[X|\sigma(Y)]$$

To prove that conditional expectations exist, we give a geometric argument, involving approximation in a Hilbert space.

Before we start the second proof, recall that a Hilbert space V is a vector space which is equipped with an inner product, which we will denote $\langle v_1, v_2 \rangle$. Now an inner product automatically induces a norm (length), and angle

$$\|v\| = \langle v, v \rangle^{\frac{1}{2}} \quad \cos \theta = \frac{\langle v_1, v_2 \rangle}{\|v_1\| \|v_2\|}$$

Here θ is the angle between v_1 and v_2 . We say that v_1, v_2 are orthogonal if $\langle v_1, v_2 \rangle = 0$. Hilbert spaces are also complete, i.e. every Cauchy sequence in V converges (to a vector in V).

Suppose that W is a complete subspace of V . We then have the notion of *orthogonal projection* onto W . Given any vector $v \in V$, there exists a decomposition

$$v = v^{\parallel} + v^{\perp}$$

a unique vector w with the following properties:

- (1) $v^{\parallel} \in W$.
- (2) $v^{\perp} \perp W$, i.e. $\langle v^{\perp}, w \rangle = 0$ for all $w \in W$.
- (3) $\|v - v^{\parallel}\| = \inf\{\|v - w\| : w \in W\}$.

Thus v^\parallel is the vector in W which is the *best approximation* of v : It lies closer to v than any other $w \in W$. v^\parallel is called the *orthogonal projection of v onto W* .

Recall also that $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ is a Hilbert space, with inner product $\langle X, Y \rangle = \mathbb{E}XY$ and induced norm $\|X\|_2 = (\mathbb{E}X^2)^{\frac{1}{2}}$.

Proof of Thm. 11.1.6: First assume that $X \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$. Note that $\mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})$ is a closed subspace of $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$, and thus there exists a decomposition

$$X = Z + Y \quad \text{where} \quad Z \in \mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P}) \quad \text{and} \quad Y \perp \mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})$$

Moreover, $\|X - Z\|_2 = \inf\{\|X - U\|_2 : U \in \mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})\}$. Now Z is clearly \mathcal{G} -measurable. Also, if $G \in \mathcal{G}$, then $I_G \in \mathcal{L}^2(\Omega, \mathcal{G})$ and so $Y \perp I_G$. Hence

$$\mathbb{E}[Z; G] = \langle Z, I_G \rangle = \langle X, I_G \rangle = \mathbb{E}[X; G] \quad \text{all } G \in \mathcal{G}$$

It follows that $Z = \mathbb{E}[X|\mathcal{G}]$ a.s.

For $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, we use an approximation argument. First assume that $X \geq 0$, and for $n \in \mathbb{N}$, define $X_n := X \wedge n$. Then $X_n \uparrow X$, and each $X_n \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$. By the above, there are $Z_n \in \mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})$ such that $Z_n = \mathbb{E}[X_n|\mathcal{G}]$ a.s. Next, note that if $n \leq m$, then $X_n \leq X_m$, and thus $Z_n \leq Z_m$ a.s.: For if $\varepsilon > 0$, and $G_\varepsilon := \{Z_n - Z_m > \varepsilon\}$, then $G_\varepsilon \in \mathcal{G}$, so that $0 \leq \varepsilon \cdot \mathbb{P}G_\varepsilon \leq \mathbb{E}[Z_n - Z_m; G_\varepsilon] = \mathbb{E}[X_n - X_m; G_\varepsilon] \leq 0$. Hence $\mathbb{P}G_\varepsilon = 0$ for all $\varepsilon > 0$. Now $\{Z_n > Z_m\} = \bigcup_{k \in \mathbb{N}} G_{\frac{1}{k}}$, and hence $\mathbb{P}(Z_n > Z_m) = 0$, i.e. $Z_n \leq Z_m$ a.s., for each pair $n \leq m$. Taking the (countable) intersection over all such pairs yields $\mathbb{P}(Z_n \text{ is an increasing sequence}) = 1$, i.e. the sequence $(Z_n)_n$ is increasing a.s. Define $Z = \limsup_n Z_n$. Then Z is \mathcal{G} -measurable, and $Z_n \uparrow Z$ a.s. If $G \in \mathcal{G}$, then by two applications of the MCT we have

$$\mathbb{E}[Z; G] = \lim_n \mathbb{E}[Z_n; G] = \lim_n \mathbb{E}[X_n; G] = \mathbb{E}[X; G]$$

Hence $Z = \mathbb{E}[X|\mathcal{G}]$ a.s.

The existence of $\mathbb{E}[X|\mathcal{G}]$ for integrable X follows by decomposition into positive and negative parts.

The a.s. uniqueness of $\mathbb{E}[X|\mathcal{G}]$ is straightforward: If Z, Z' are two versions of $\mathbb{E}[X|\mathcal{G}]$, and $\varepsilon > 0$ then $G_\varepsilon := \{Z - Z' > \varepsilon\} \in \mathcal{G}$, and hence $0 \leq \varepsilon \cdot \mathbb{P}G_\varepsilon \leq \mathbb{E}[Z - Z'; G_\varepsilon] = \mathbb{E}[X - X; G_\varepsilon] = 0$. Arguing as above, we see that $\mathbb{P}(Z > Z') = 0$. By symmetry, $\mathbb{P}(Z' > Z) = 0$ as well, i.e. $Z = Z'$ a.s.

Theorem 11.1.8 (Properties of Conditional Expectation)

The following are true for random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ whenever the expressions occurring inside a conditional expectation are integrable.

- (a) $\mathbb{E}[\mathbb{E}[X|\mathcal{G}]] = \mathbb{E}X$ a.s.;
- (b) If X is \mathcal{G} -measurable, then $\mathbb{E}[X|\mathcal{G}] = X$ a.s.
- (c) **LINEARITY:** $\mathbb{E}[a_1X_1 + a_2X_2|\mathcal{G}] = a_1\mathbb{E}[X_1|\mathcal{G}] + a_2\mathbb{E}[X_2|\mathcal{G}]$ a.s.
- (d) **POSITIVITY:** If $X \geq 0$, then $\mathbb{E}[X|\mathcal{G}] \geq 0$ a.s.
- (e) **cMCT:** If $0 \leq X_n \uparrow X$, then $\mathbb{E}[X_n|\mathcal{G}] \uparrow \mathbb{E}[X|\mathcal{G}]$ a.s.
- (f) **cFATOU:** If $X_n \geq 0$, then $\mathbb{E}[\liminf_n X_n|\mathcal{G}] \leq \liminf_n \mathbb{E}[X_n|\mathcal{G}]$ a.s.
- (g) **cDCT:** If $|X_n| < Y$ (all $n \in \mathbb{N}$) for some integrable Y , and if $X_n \rightarrow X$, then $\mathbb{E}[X_n|\mathcal{G}] \rightarrow \mathbb{E}[X|\mathcal{G}]$ a.s.
- (h) **PROJECTION:** $\mathbb{E}[X \cdot \mathbb{E}[Y|\mathcal{G}]] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}] \cdot Y] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}] \cdot \mathbb{E}[Y|\mathcal{G}]]$.
- (i) If Y is \mathcal{G} -measurable, then $\mathbb{E}[YX|\mathcal{G}] = Y\mathbb{E}[X|\mathcal{G}]$ a.s.
- (j) **TOWER:** If $\mathcal{H} \subseteq \mathcal{G}$, then $\mathbb{E}[\mathbb{E}[X|\mathcal{G}|\mathcal{H}]] = \mathbb{E}[\mathbb{E}[X|\mathcal{H}]] = \mathbb{E}[X|\mathcal{H}]$.
- (k) **INDEPENDENCE:** If \mathcal{H} is independent of $\sigma(X) \vee \mathcal{G}$, then $\mathbb{E}[X|\mathcal{G} \vee \mathcal{H}] = \mathbb{E}[X|\mathcal{G}]$ a.s.

Exercise 11.1.9 Prove Thm. 11.1.8(a), (b), (c), (d).

[Hint: (c) means that if Y_k are versions of $\mathbb{E}[X_k|\mathcal{G}]$ for $k = 1, 2$, then $a_1Y_1 + a_2Y_2$ is a version of $\mathbb{E}[a_1X_1 + a_2X_2|\mathcal{G}]$.

For (d), let $Z = \mathbb{E}[X|\mathcal{G}]$ a.s., and note that $\{Z < 0\} = \bigcup_{n \in \mathbb{N}} \{Z < -\frac{1}{n}\} \in \mathcal{G}$.

□

Proof of Thm. 11.1.8(e)–(k):

(e): Suppose that $0 \leq X_n \uparrow X$, and define $Y_n := \mathbb{E}[X_n|\mathcal{G}]$ a.s. By (d), Y_n is increasing a.s. Define $Y = \limsup_n Y_n$, so that Y is \mathcal{G} -measurable and $Y_n \uparrow Y$ a.s. If $G \in \mathcal{G}$, then by the MCT, $\mathbb{E}[X; G] = \lim_n \mathbb{E}[X_n; G] = \lim_n \mathbb{E}[Y_n; G] = \mathbb{E}[Y; G]$. Hence $Y = \mathbb{E}[X|\mathcal{G}]$ a.s.

(f): Let $Z_n := \inf_{k \geq n} X_k$ a.s., so that $Z_n \uparrow \liminf_n X_n$. Since $Z_n \leq X_k$ whenever $k \geq n$, we have $\mathbb{E}[Z_n|\mathcal{G}] \leq \mathbb{E}[X_k|\mathcal{G}]$ a.s. whenever $k \geq n$, and hence $\mathbb{E}[Z_n|\mathcal{G}] \leq \inf_{k \geq n} \mathbb{E}[X_k|\mathcal{G}]$ a.s. Now by cMCT,

$$\mathbb{E}[\liminf_n X_n|\mathcal{G}] = \lim_n \mathbb{E}[Z_n|\mathcal{G}] \leq \lim_n \inf_{k \geq n} \mathbb{E}[X_k|\mathcal{G}] = \liminf_n \mathbb{E}[X_n|\mathcal{G}]$$

(g): $Y \pm X_n$ are non-negative random variables, so by cFATOU,

$$\mathbb{E}[Y|\mathcal{G}] + \liminf_n (\pm \mathbb{E}[X_n|\mathcal{G}]) = \liminf_n \mathbb{E}[Y \pm X_n|\mathcal{G}] \geq \mathbb{E}[\liminf_n Y \pm X_n|\mathcal{G}] = \mathbb{E}[Y|\mathcal{G}] \pm \mathbb{E}[X|\mathcal{G}] \quad \text{a.s.}$$

Since $\mathbb{E}[Y|\mathcal{G}]$ is integrable, it is finite a.s., and hence can be cancelled to yield $\liminf_n (\pm \mathbb{E}[X_n|\mathcal{G}] \geq \pm \mathbb{E}[X|\mathcal{G}])$, which implies

$$\mathbb{E}[X|\mathcal{G}] \leq \liminf_n \mathbb{E}[X|\mathcal{G}] \leq \limsup_n \mathbb{E}[X_n|\mathcal{G}] \leq \mathbb{E}[X|\mathcal{G}] \quad \text{a.s.}$$

(h): This follows from the usual properties of *projections* if $X, Y \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$: For suppose that $X = X_{\parallel} + X_{\perp}, Y = Y_{\parallel} + Y_{\perp}$ are decompositions of X, Y into components parallel and perpendicular to $\mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})$, so that $X_{\parallel} = \mathbb{E}[X|\mathcal{G}], Y_{\parallel} = \mathbb{E}[Y|\mathcal{G}]$ (by the second proof of Thm. 11.1.6). Then

$$\mathbb{E}[X \cdot \mathbb{E}[Y|\mathcal{G}]] = \langle X, Y_{\parallel} \rangle = \langle X_{\parallel}, Y_{\parallel} \rangle = \mathbb{E}[\mathbb{E}[X|\mathcal{G}] \cdot \mathbb{E}[Y|\mathcal{G}]] \quad \text{a.s.}$$

because $\langle X_{\perp}, Y_{\parallel} \rangle = 0$. If $X, Y \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ are non-negative, we may define $X_n := X \wedge n, Y_n := Y \wedge n$. Then $0 \leq X_n \uparrow X$ and $0 \leq Y_n \uparrow Y$, and $X_n, Y_n \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$. It follows by the MCT and cMCT that

$$\mathbb{E}[X \cdot \mathbb{E}[Y|\mathcal{G}]] = \lim_n \mathbb{E}[X_n \cdot \mathbb{E}[Y_n|\mathcal{G}]] = \lim_n \mathbb{E}[\mathbb{E}[X_n|\mathcal{G}] \cdot \mathbb{E}[Y_n|\mathcal{G}]] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}] \cdot \mathbb{E}[Y|\mathcal{G}]] \quad \text{a.s.}$$

(i): If $Y = I_G$ is an indicator function, and $G' \in \mathcal{G}$, then

$$\mathbb{E}[Y\mathbb{E}[X|\mathcal{G}]; G'] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]I_G I_{G'}] = \mathbb{E}[X; G \cap G'] = \mathbb{E}[YX; G']$$

Hence $Y\mathbb{E}[X|\mathcal{G}]$ is a version of $\mathbb{E}[YX|\mathcal{G}]$. The result now follows by linearity and cMCT.

(j): Consider the case where $X \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$. Since $\mathcal{L}^2(\Omega, \mathcal{H}, \mathbb{P}) \subseteq \mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P}) \subseteq \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ are closed Hilbert subspaces, the result follows from the fact that a projection of a projection is a projection. Alternatively, let

$$Y := \mathbb{E}[X|\mathcal{G}] \quad \text{a.s.} \quad Z := \mathbb{E}[\mathbb{E}[X|\mathcal{G}]|\mathcal{H}] = \mathbb{E}[Y|\mathcal{H}] \quad \text{a.s.}$$

If $H \in \mathcal{H} \subseteq \mathcal{G}$, then

$$\mathbb{E}[Z; H] = \mathbb{E}[Y; H] = \mathbb{E}[X; H]$$

and hence Z is a version of $\mathbb{E}[X|\mathcal{H}]$, i.e. $\mathbb{E}[\mathbb{E}[X|\mathcal{G}]|\mathcal{H}] = \mathbb{E}[X|\mathcal{H}]$ a.s.

The fact that $\mathbb{E}[\mathbb{E}[X|\mathcal{H}]|\mathcal{G}] = \mathbb{E}[X|\mathcal{H}]$ a.s. follows directly (i).

(k): Let $Y := \mathbb{E}[X|\mathcal{G}]$. Since Y is certainly $\mathcal{G} \vee \mathcal{H}$ -measurable, we must show that $\mathbb{E}[Y; F] = \mathbb{E}[X; F]$ for all $F \in \mathcal{G} \vee \mathcal{H}$. Now let $\mathcal{C} := \{G \cap H : G \in \mathcal{G}, H \in \mathcal{H}\}$, and let $\mathcal{D} := \{F \in \mathcal{G} \vee \mathcal{H} : \mathbb{E}[Y; F] = \mathbb{E}[X; F]\}$. First note that $\mathcal{C} \subseteq \mathcal{D}$: For if $G \in \mathcal{G}, H \in \mathcal{H}$, then $\mathbb{E}[X; G \cap H] = \mathbb{E}[XI_G]\mathbb{E}[I_H]$, by independence, and so

$$\mathbb{E}[X; G \cap H] = \mathbb{E}[X; G]\mathbb{E}[I_H] = \mathbb{E}[Y; G]\mathbb{E}[I_H] = \mathbb{E}[Y; G \cap H]$$

since YI_G is independent of \mathcal{H} . It is straightforward to verify that \mathcal{C} is a π -system that generates $\mathcal{G} \vee \mathcal{H}$, and that \mathcal{D} is a λ -system. Hence by the π - λ Theorem, $\mathcal{D} = \mathcal{G} \vee \mathcal{H}$.

—

Definition 11.1.10 Let U be an open subset of \mathbb{R}^n . A function $g : U \rightarrow \mathbb{R}$ is said to be **convex** if and only if for any $x, y \in U$ and any $\lambda \in [0, 1]$ we have

$$g(\lambda x + (1 - \lambda)y) \leq \lambda g(x) + (1 - \lambda)g(y)$$

Remarks 11.1.11 The following remarks feature in the proof of Jensen's inequality, the next proposition, and should be digested thoroughly.

- (a) Recall that if x, y are points in \mathbb{R}^n , then $\{\lambda x + (1 - \lambda)y : 0 \leq \lambda \leq 1\}$ is simply the line segment in \mathbb{R}^n joining x to y . The point with coordinates

$$(\lambda x + (1 - \lambda)y, g(\lambda x + (1 - \lambda)y))$$

is simply a point on the graph of g between x and y . On the other hand, the point

$$(\lambda x + (1 - \lambda)y, \lambda g(x) + (1 - \lambda)g(y))$$

is a point on the line segment joining $(y, g(y))$ to $(x, g(x))$. These two points have the same x -coordinate, namely $\lambda x + (1 - \lambda)y$. We can now interpret convexity geometrically: A function g is convex if and only if its graph lies below any chord (line segment) joining two points on the graph of g .

- (b) In particular, if $g : \mathbb{R} \rightarrow \mathbb{R}$ has $g'' \geq 0$, then g is a convex function. The functions x^+ , x^- , $|x|$ are also convex.
- (c) Let $g : U \rightarrow \mathbb{R}$, where U is an open subinterval of \mathbb{R} . For $u, v \in U$, define $\Delta(u, v) = \frac{g(u) - g(v)}{u - v}$. Geometrically, $\Delta(u, v)$ is the slope of the chord joining $(u, g(u))$ to $(v, g(v))$ on the graph of g . Then g is convex if and only if $u < v < w$ in U implies $\Delta(u, v) \leq \Delta(v, w)$. This is easy to see geometrically. A more rigorous proof: If $u < v < w$, define $\lambda = \frac{v - w}{u - w}$. Then $v = \lambda u + (1 - \lambda)w$. It follows that

$$g(v) \leq \lambda g(u) + (1 - \lambda)g(w) = \frac{v - w}{u - w}g(u) + \frac{u - v}{u - w}g(w)$$

Hence

$$(u - v)g(v) + (v - w)g(v) \geq (v - w)g(u) + (u - v)g(w)$$

Rearranging yields the result.

- (d) Now if $v < w$ in U and we let $u \uparrow v$, then

- (i) $\Delta(u, v)$ increases as $u \uparrow v$.
- (ii) $\Delta(u, v) \leq \Delta(v, w)$, and thus $\Delta(u, v)$ is bounded from above as $u \uparrow v$.

Since a sequence which is increasing and bounded from above must converge, $D^-(v) = \lim_{u \uparrow v} \Delta(u, v)$ exists. Similar reasoning shows that $D^+(v) = \lim_{w \downarrow v} \Delta(v, w)$ must exist for every $v \in U$. Thus left- and right derivatives exist at every point v . Moreover, $D^-(v) \leq D^+(v)$, because each $\Delta(u, v)$ is \leq each $\Delta(v, w)$. If these limits are equal, then g is differentiable at v .

- (e) A convex function is automatically continuous, and thus a Borel function: For let $v \in U$. If there is a discontinuity at v , then it is easy to see that either $\lim_{u \uparrow v} \Delta(u, v)$ or $\lim_{w \downarrow v} \Delta(v, w)$ does not exist.

□

Proposition 11.1.12 (Jensen's inequality)

Suppose that $g : U \rightarrow \mathbb{R}$ is a convex function on an open interval $U \subseteq \mathbb{R}$, and that X is a random variable with values in U (a.s.) such that both X and $g(X)$ have finite expected values. Then

$$\mathbb{E}[g(X)|\mathcal{G}] \geq g(\mathbb{E}[X|\mathcal{G}])$$

Proof: We use notation and results from Remarks 6.2.5. Let $v \in U$, and let $D^-(v) = \lim_{u \uparrow v} \Delta(u, v)$ and $D^+(v) = \lim_{w \downarrow v} \Delta(v, w)$. Then $D^-(v), D^+(v)$ both exist, and $D^-(v) \leq D^+(v)$. Now suppose that m is a real number satisfying $D^-(v) \leq m \leq D^+(v)$, and that $x \in U$. We consider two cases: If (i) $x \leq v$, then $\Delta(x, v) \leq D^-(v)$ (since $\Delta(u, v)$ increases as $u \uparrow v$) and thus $\Delta(x, v) \leq m$. It follows that $g(x) \geq m(x - v) + g(v)$. Next, if (ii) $x \geq v$, then $\Delta(v, x) \geq D^+(v)$ (because $\Delta(v, w)$ decreases as $w \downarrow v$) and thus $\Delta(v, x) \geq m$. It follows that $g(x) \geq m(x - v) + g(v)$. Hence, in either case, we have

$$g(x) \geq m(x - v) + g(v)$$

for any $v \in U$, any $x \in U$, and any $D^-(v) \leq m \leq D^+(v)$.

We are now ready to prove Jensen's inequality: Put $v = \mathbb{E}[X|\mathcal{G}]$. Then

$$g(X) \geq m(X - \mathbb{E}[X|\mathcal{G}]) + g(\mathbb{E}[X|\mathcal{G}]) \text{ a.s.} \quad \text{whenever } D^-(\mathbb{E}[X|\mathcal{G}]) \leq m \leq D^+(\mathbb{E}[X|\mathcal{G}])$$

If we now take conditional expectations on both sides, then

$$\mathbb{E}[g(X)|\mathcal{G}] \geq m(\mathbb{E}[X|\mathcal{G}] - \mathbb{E}[X|\mathcal{G}]) + \mathbb{E}[g(\mathbb{E}[X|\mathcal{G}])|\mathcal{G}] = g(\mathbb{E}[X|\mathcal{G}])$$

□

Some notation: we define

$$\mathbb{E}[X|\mathcal{G}|\mathcal{H}] := \mathbb{E}[\mathbb{E}[X|\mathcal{G}]\mathcal{H}]$$

As always, in the discrete world everything is simple:

Suppose that X is a random variable on a discrete probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and that $\mathcal{G} \subseteq \mathcal{F}$ is generated by the partition G_1, \dots, G_n . Then $\mathbb{E}(X|\mathcal{G})$ is constant on each G_i , and

$$\mathbb{E}(X|\mathcal{G})(\omega) = \frac{\mathbb{E}(X; G_i)}{\mathbb{P}(G_i)} = \mathbb{E}[X|G_i] \quad \text{for } \omega \in G_i$$

You should verify the statement in the box.

Exercise 11.1.13 Toss three fair coins one after the other, a R1 coin, an R2 coin and an R3 coin. You get to keep the coins which land H . The sample space is $\Omega = \{HHH, \dots, TTT\}$, and $\mathbb{P}(\{\omega\}) = \frac{1}{8}$ for all $\omega \in \Omega$. Let \mathcal{F}_n denote the information after the n^{th} toss, for $n = 0, 1, 2, 3$. Let X be your winnings. Calculate the random variables $\mathbb{E}[X|\mathcal{F}_n]$.

□

We end this chapter with some simple examples:

- Examples 11.1.14** (a) Suppose that X is a random variable in $\mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$, where $p \geq 1$, and let Y be a version of $\mathbb{E}(X|\mathcal{G})$. Then Y is in $\mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ as well. This is because $g(x) = |x|^p$ is convex. By Jensen's inequality, we therefore have $|\mathbb{E}(X|\mathcal{G})|^p \leq \mathbb{E}(|X|^p|\mathcal{G})$ a.s., i.e. $|Y|^p \leq \mathbb{E}(|X|^p|\mathcal{G})$. It follows that $\mathbb{E}|Y|^p \leq \mathbb{E}[\mathbb{E}(|X|^p|\mathcal{G})] = \mathbb{E}|X|^p < +\infty$, by (I).
- (b) Suppose that $X \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ (i.e. that $\text{var}(X)$ exists). If Y is a version of $\mathbb{E}(X^2|\mathcal{G})$, then $Y \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ as well, by (a), and $\mathbb{E}Y^2 \leq \mathbb{E}X^2$. Since $\mathbb{E}Y = \mathbb{E}X$ (by (I)), we thus have

$$\text{var}(Y) = \mathbb{E}Y^2 - (\mathbb{E}Y)^2 \leq \mathbb{E}X^2 - (\mathbb{E}X)^2 = \text{var}(X)$$

Thus $\text{var}(Y) \leq \text{var}(X)$. This reflects the fact that Y , being cruder, can't vary as much as X can.

□

11.2 Theory of Martingales in Discrete Time

Martingales are amongst the most important objects in probability theory, and an entire sub-discipline of finance is based on them. Brownian motion is the most important continuous-parameter martingale, and is heavily used in financial modelling. In this chapter we first introduce the basic results about discrete-time martingales at a leisurely rate, taking time to build up intuition and facility with martingale calculations. In the next chapter, we will tackle continuous-parameter martingales.

11.2.1 Stochastic Processes and Filtrations

Informally, a (discrete-parameter) stochastic process X is a family of random variables indexed by a discrete time set, i.e. $X = X_1, X_2, X_3, \dots$. The idea is that these model the outcomes of a series of random phenomena, such as the closing values of the S&P500. The X_n are thus successive values of some quantity under consideration. Note that the times of the random variables may not be evenly distributed in physical time; for example, the share index is recorded only on trading days.

We assume that the stochastic process $X = (X_n : n \in I)$ is defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The time index set I will usually be the set of natural numbers, or the set of non-negative reals, or some finite initial segment of these. For a particular outcome $\omega \in \Omega$, the sequence $X_1(\omega), X_2(\omega), \dots$ is called a *sample path* of the process. Note that *one* outcome/state-of-the-world ω determines the values of all the X_n . We only know the value of X_n at time n , and so as time n increases, so does our knowledge of the state of the world. Since information is organised in σ -algebras, we associate with each time n a σ -algebra \mathcal{F}_n modelling the knowledge at time n . We also assume that no information is lost or forgotten, so that information available at time n is also available at a later time $m > n$. This simply means that $\mathcal{F}_m \supseteq \mathcal{F}_n$. We thus model the flow of information as follows:

Definition 11.2.1 Suppose that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. An increasing sequence

$$\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_n \subseteq \cdots \subseteq \mathcal{F}$$

of σ -algebras on Ω is called a **filtration**. We shall always assume that \mathcal{F}_0 contains all the sets of measure 0.

We also define

$$\mathcal{F}_\infty = \sigma\left(\bigcup_n \mathcal{F}_n\right) \subseteq \mathcal{F}$$

\mathcal{F}_n represents the available information at time n , i.e. it contains all events A for which it is possible to decide at time n whether A has occurred or not.

Suppose that S_t is the share price at time t . We know S_2 at time $t = 2$. Thus each of the following events can be decided at time $t = 2$: Whether or not $X_2 = 5.00$; whether or not X_2 lies between 13.50 and 15.76, etc. It therefore follows that X_2 must be \mathcal{F}_2 -measurable, i.e. that $\sigma(X_2) \subseteq \mathcal{F}_2$. Moreover, X_1 is also known at $t = 2$, so $\sigma(X_1, X_2) \subseteq \mathcal{F}_2$. However, at $t = 2$ we do not know the share price at time $t = 3$. Thus X_3 is not \mathcal{F}_2 -measurable, although it is, of course \mathcal{F}_3 -measurable.

In essence, to model the fact that the value of X_m is known at a later time n , we need to add the restriction that X_m is \mathcal{F}_n -measurable for all $n \geq m$. This just means that $\sigma(X_1, \dots, X_n) \subseteq \mathcal{F}_n$, and so we define:

Definition 11.2.2 A stochastic process $X = (X_n, n \in I)$ is said to be *adapted* to a filtration $\mathcal{F}_n, n \in I$ provided that each X_n is \mathcal{F}_n -measurable. It follows trivially that this is the case if and only if

$$\sigma(X_1, \dots, X_n) \subseteq \mathcal{F}_n$$

Exercise 11.2.3 Make sure that you can prove this trivial result.

□

Note that to say that X is adapted to \mathcal{F}_n simply means that the random variables X_n do not contain more information than the \mathcal{F}_n , although they may contain strictly less.

Note also that $\mathcal{F}_n = \sigma(X_1, \dots, X_n)$ is the smallest filtration with respect to which X is adapted, i.e. that if X is also adapted to a filtration \mathcal{G}_n , then $\mathcal{F}_n \subseteq \mathcal{G}_n$. The filtration \mathcal{F}_n contains just the information in the values of X up to time n , and is called the *natural* or *canonical* filtration of X . It contains just as much information as is contained in the X_n , and no more.

11.2.2 Martingales, Submartingales, Supermartingales

Martingales model a fair game, submartingales a favourable game, and supermartingales an unfavourable game. Here is the definition:

Definition 11.2.4 A stochastic process $X = (X_n : n \in \mathbb{N})$ is called a **supermartingale** (respectively **submartingale**) with respect to a filtration $\mathcal{F}_n, n \in \mathbb{N}$ if and only if

- (a) Each $X_n \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})^a$
- (b) X is adapted to $\mathcal{F}_n, n \in \mathbb{N}$.
- (c) $\mathbb{E}[X_{n+1}|\mathcal{F}_n] \leq X_n$ (respectively, $\mathbb{E}[X_{n+1}|\mathcal{F}_n] \geq X_n$) for each $n \in \mathbb{N}$

A **martingale** is simultaneously a sub- and a supermartingale, i.e. it satisfies $\mathbb{E}[X_{n+1}|\mathcal{F}_n] = X_n$ for each $n \in \mathbb{N}$.

When we say that X is a (super/sub-)martingale, but we don't mention a specific filtration, then the natural filtration should be used.

^ai.e. each X_n is *integrable*, which just means that $\mathbb{E}X_n$ exists, and is finite.

(Note that we've taken \mathbb{N} as the index set. You shouldn't have any trouble generalizing the definition to the case where the index set is some initial segment $\{0, 1, 2, \dots, T\}$ of \mathbb{N}).

Think of X_n as your total fortune after the n^{th} round of a gambling game. If X is a supermartingale, your expected fortune at time $n+1$ is less than your fortune at time n . It follows that this particular game is unfavourable, i.e. that you are likely to lose. If X is a martingale, then your expected fortune equals your present fortune: You are just as likely to win as to lose, and the game is fair.

Examples 11.2.5 (a) Suppose that the $X_n, n \in \mathbb{N}$ are independent random variables with $\mathbb{E}X_n = 0$, and that $\mathcal{F}_n, n \in \mathbb{N}$ is the natural filtration. Define $S_n = X_1 + \dots + X_n$. Clearly $S = (S_n : n \in \mathbb{N})$ is a stochastic process adapted to $\mathcal{F}_n, n \in \mathbb{N}$, and each S_n is integrable. Moreover,

$$\mathbb{E}[S_{n+1}|\mathcal{F}_n] = \mathbb{E}[X_1|\mathcal{F}_n] + \dots + \mathbb{E}[X_n|\mathcal{F}_n] + \mathbb{E}[X_{n+1}|\mathcal{F}_n]$$

Since X_m is \mathcal{F}_n -measurable for $m \leq n$, it follows that $\mathbb{E}[X_m|\mathcal{F}_n] = X_m$ if $m \leq n$. Moreover, since the X_m are independent random variables, X_{n+1} is independent of \mathcal{F}_n , and thus we have $\mathbb{E}[X_{n+1}|\mathcal{F}_n] = \mathbb{E}X_{n+1} = 0$. Hence

$$\mathbb{E}[S_{n+1}|\mathcal{F}_n] = X_1 + \dots + X_n + 0 = S_n$$

which proves that $S_n, n \in \mathbb{N}$ is a martingale.

- (b) If we have the same situation as in (a), but with $\mathbb{E}X_n > 0$ for all n , then $S_n, n \in \mathbb{N}$ is a submartingale.
- (c) If $X_n, n \in \mathbb{N}$ are random variables with the same mean $\mu = 0$ and the same variance σ^2 , and if $S_n = X_1 + X_2 + \dots + X_n$, then the process $W_n = S_n^2 - n\sigma^2$ is a martingale with respect to the natural filtration of the X_n . First note that each W_n is integrable if and only if S_n^2 is, but this follows because the variances $\sigma^2 = \mathbb{E}X_n^2$ exist, so that each $X_n \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$. To verify the martingale property, observe that $S_{n+1}^2 = S_n^2 + 2S_nX_{n+1} + X_{n+1}^2$. Further observe that $\mathbb{E}[S_nX_{n+1}|\mathcal{F}_n] = S_n\mathbb{E}[X_{n+1}|\mathcal{F}_n]$, because S_n is \mathcal{F}_n -measurable, and that

$\mathbb{E}[X_{n+1}|\mathcal{F}_n] = \mathbb{E}X_{n+1} = 0$, since X_{n+1} is independent of \mathcal{F}_n . Thus:

$$\begin{aligned}\mathbb{E}[W_{n+1} - W_n|\mathcal{F}_n] &= \mathbb{E}[(S_n + X_{n+1})^2 - S_n^2 - \sigma^2|\mathcal{F}_n] \\ &= 2\mathbb{E}[S_n X_{n+1}|\mathcal{F}_n] + \mathbb{E}[X_{n+1}^2|\mathcal{F}_n] - \sigma^2 \\ &= 2S_n \mathbb{E}[X_{n+1}|\mathcal{F}_n] + \mathbb{E}X_{n+1}^2 - \sigma^2 \\ &= 2S_n \cdot \mathbb{E}X_{n+1} + \text{Var}(X_n) - \sigma^2 \\ &= 0\end{aligned}$$

- (d) Suppose that X_n are non-negative random variables with $\mathbb{E}X_n = 1$. Put $M_0 = 1$, and define

$$M_n = X_1 \cdot X_2 \cdots X_n$$

for $n > 0$. Assume that each M_n is integrable. It is left as an exercise to show that M_n is a martingale.

- (e) Consider a random walk. If it is symmetric, it is a martingale. If the probability p of going up is < 0.5 , it is a supermartingale.
- (f) One more interesting martingale demonstrates the accumulation of information about the value of a random variable over time. Let Y be an integrable random variable (i.e. \mathcal{F} -measurable). We do not necessarily know the value of Y at time n — there may not be enough information available. However, as time passes, we expect that our estimate will become more accurate. At time n , the best available approximation to Y is $Y_n = \mathbb{E}[Y|\mathcal{F}_n]$. We now show that Y_n is a martingale (with respect to the natural filtration). Firstly,

$$\mathbb{E}Y_n = \mathbb{E}[\mathbb{E}(Y|\mathcal{F}_n)] = \mathbb{E}Y$$

by the “Tower Property”. This shows that each Y_n is integrable if Y is. Next,

$$\mathbb{E}[Y_{n+1}|\mathcal{F}_n] = \mathbb{E}[\mathbb{E}[Y|\mathcal{F}_{n+1}]|\mathcal{F}_n] = \mathbb{E}[Y|\mathcal{F}_n] = Y_n$$

by the Tower Property again. This proves the result.

What this means is that there are no trends in our estimates of Y . At each new time step, our revised estimate is just as likely to go up as it is to go down, and is expected to remain at the same value as our previous estimate. This makes sense: If we *expected* our estimates to increase, for example, then our estimates would not have been the best available. We ought to have built the expectation of increase into our estimates already.

- (g) Note that if X_n is a martingale, and if φ is a *convex* function, then $\varphi(X_n)$ is a submartingale. Indeed,

$$\mathbb{E}[\varphi(X_{n+1})|\mathcal{F}_n] \geq \varphi(\mathbb{E}[X_{n+1}|\mathcal{F}_n]) = \varphi(X_n)$$

by Jensen’s inequality. It follows that if X_n is a martingale and if $p > 1$, then $|X_n|^p$ is a submartingale.

□

Remarks 11.2.6 (a) If $X_n, n \in \mathbb{N}$ is a martingale, then $\mathbb{E}X_n = \mathbb{E}X_0$ for all n , i.e. all the X_n have the same mean. This is an easy exercise.

- (b) We have defined the martingale property with respect to a filtration. Thus if X_n is a martingale with respect to one filtration, it may not be with respect to another. However, if X_n is a martingale with respect to some filtration \mathcal{G}_n , and if $\mathcal{F}_n = \sigma(X_1, \dots, X_n)$ is the natural filtration, then X_n is also a martingale with respect to \mathcal{F}_n . To see this, first note that each X_n is \mathcal{G}_n -measurable (because X_n is adapted to \mathcal{G}_n — part of the definition of *martingale*). Thus $\mathcal{F}_n \subseteq \mathcal{G}_n$ for each n . It now follows by the Tower Property that

$$\mathbb{E}[X_{n+1}|\mathcal{F}_n] = \mathbb{E}[\mathbb{E}[X_{n+1}|\mathcal{G}_n]|\mathcal{F}_n] = \mathbb{E}[X_n|\mathcal{F}_n] = X_n$$

The last equality holds by “Taking out what is known”, because X_n is \mathcal{F}_n -measurable. It is now not hard to see that if X_n is a martingale with respect to one filtration, it will also be a martingale with respect to any poorer (in information) filtration to which it is adapted.

- (c) The converse of (b) is not true: If X_n is a martingale with respect to the natural filtration, it may not be a martingale with respect to a richer (in information) filtration. Find a simple example!
- (d) Note that if X_n is a martingale, and if $m \geq n$, then $\mathbb{E}[X_m|\mathcal{F}_n] = X_n$. This is left as another exercise in the use of the Tower Property.

□

The following exercise will prove extremely useful:

Exercise 11.2.7 (Orthogonality of Martingale Increments)

Prove that if M_n is a martingale, then

$$\mathbb{E}[(M_n - M_m)^2|\mathcal{F}_k] = \mathbb{E}[M_n^2 - M_m^2|\mathcal{F}_k] \quad k \leq m \leq n$$

Deduce that

$$\mathbb{E}[M_n]^2 = \mathbb{E}M_0^2 + \sum_{m=1}^n \mathbb{E}[(M_m - M_{m-1})^2]$$

□

11.2.3 Games and Strategies

Suppose that you take part in a game of chance, e.g. a game of coin tossing, roulette, or investing in the stock market. The game is repeated many times, and you place a bet each time. Let $\xi_n, n \in \mathbb{N}$ be a sequence of integrable random variables which represent your winnings (or losses, if negative) per unit stake in the n^{th} game. Thus, if you had wagered a stake C_n on the n^{th} game, you would have won $C_n \xi_n$.

If you played unit stakes all the way through, your total winnings after the n^{th} game would be

$$S_n = \xi_1 + \dots + \xi_n \quad \text{for } n \geq 1$$

Note that $S_0 = 0$, because you haven't won or lost anything yet.

If the game is *fair* then your chance of winning is the same as your chance of losing, and thus $\mathbb{E}\xi_n = 0$. In that case, S_n is a martingale with respect to the natural filtration

$\mathcal{F}_n = \sigma(\xi_1, \dots, \xi_n) = \sigma(S_1, \dots, S_n)$. Similarly, if the game is unfavourable to you, then at time n you expect your winnings at time $n + 1$ to be less than your current winnings, i.e. $\mathbb{E}[S_{n+1}|\mathcal{F}_n] \leq S_n$. Thus an unfavourable game is modelled by a supermartingale. A favourable game will clearly be modelled by a submartingale.

Suppose now that you have a *system*, i.e. a gambling strategy, which tells you when to bet, how much to bet etc. Your system, call it C , will tell you what stake C_n you should place on the n^{th} game. We allow negative stakes as well (which are essentially bets that you will lose)². In that case, your total winnings after the n^{th} game will be

$$W_n = C_1\xi_1 + \dots + C_n\xi_n$$

Now note that $\xi_n = S_n - S_{n-1} = \Delta S_n$, and thus that

$$W_n = \sum_{k=1}^n C_k(S_k - S_{k-1}) = \sum_{k=1}^n C_k \Delta S_k$$

which looks like a Riemann–Stieltjes sum³. Your strategy $C = (C_n : n \in I)$ is also a stochastic process, but since we have to decide what stake to wager before the outcome of the n^{th} game is known, we must be able to decide the value of C_n on the basis of information available at time $n - 1$ (i.e. after the $(n - 1)^{\text{th}}$ game). Thus each C_n is \mathcal{F}_{n-1} –measurable. We have a name for this:

Definition 11.2.8 A stochastic process C is called *previsible* (or *non-anticipative*, or *predictable*), with respect to a filtration \mathcal{F}_n provided that each C_n is \mathcal{F}_{n-1} –measurable, for $n \geq 1$. Note that C_0 is not defined.

Thus a gambling strategy is just a previsible process.

Consider an arbitrary adapted stochastic process Y_n . Then in general Y_n may exhibit both purely random behaviour and long-term trends. For example, for supermartingales the long-term trend is that it tends to decrease. Purely random behaviour is described by martingales, and trends are known beforehand, i.e. are previsible. We thus attempt to decompose Y_t into a martingale part and a previsible part, i.e. we try to write

$$Y_n = M_n + A_n$$

where M_n is a martingale, with $M_0 = Y_0$, and A_n is previsible, with $A_0 = 0$. In engineering, A_n is called the *signal*, and M_n the *noise*.

Suppose that we can actually find such a decomposition. We would then have

$$Y_{n+1} - Y_n = (M_{n+1} - M_n) + (A_{n+1} - A_n)$$

Taking conditional expectations immediately yields

$$A_{n+1} - A_n = \mathbb{E}[Y_{n+1}|\mathcal{F}_n] - Y_n$$

so that

$$M_{n+1} - M_n = Y_{n+1} - \mathbb{E}[Y_{n+1}|\mathcal{F}_n]$$

We now use this pair of equations to *define* M_n and A_n in the next theorem.

²We need negative stakes to model short sales, which are essentially just bets that a stock will lose value

³The Riemann–Stieltjes integral is discussed in Chapter ??

Theorem 11.2.9 (Doob Decomposition Theorem)

Every process Y_n has a unique decomposition

$$Y_n = M_n + A_n$$

where M_n is a martingale with $M_0 = Y_0$, and A_n is previsible, null at $n = 0$. Moreover, if Y_n is a supermartingale, then A_n is decreasing.

Proof: Define M_n, A_n inductively by

$$\begin{cases} M_0 = Y_0, & M_{n+1} = M_n + Y_{n+1} - \mathbb{E}[Y_{n+1}|\mathcal{F}_n] \\ A_0 = 0, & A_{n+1} = A_n + \mathbb{E}[Y_{n+1}|\mathcal{F}_n] - Y_n \end{cases}$$

It is clear that M_n is a martingale and that A_n is previsible. Moreover

$$Y_m - Y_{m-1} = (M_m - M_{m-1}) + (A_m - A_{m-1})$$

summing over m from $m = 1$ to $m = n$ yields

$$Y_n = M_n + A_n$$

as required.

To see that this decomposition is unique, suppose that $Y_n = M'_n + A'_n$ is another decomposition with the same properties. We show by induction on n that $M = M', A = A'$: Note that $M_0 = M'_0$ by definition. Suppose that $M_n = M'_n$, and, consequently, that $A_n = A'_n$. Then

$$M_{n+1} - M'_{n+1} = A_{n+1} - A'_{n+1}$$

Taking conditional expectations with respect to \mathcal{F}_n , we obtain

$$0 = M_n - M'_n = A'_{n+1} - A_{n+1}$$

because A, A' are previsible. Hence $A_{n+1} = A'_{n+1}$, and so $M_{n+1} = M'_{n+1}$ as well. By induction, we have $M_n = M'_n, A_n = A'_n$ for all $n \in \mathbb{N}$. This proves that the Doob Decomposition is unique.

If X_n is a supermartingale, then $\mathbb{E}[X_{n+1}|\mathcal{F}_n] \leq X_n$, so the definition of A_{n+1} implies that $A_{n+1} \leq A_n$.

◄

Exercise 11.2.10 Suppose that Y_t is a martingale. By Jensen's inequality, Y_t^2 will be a submartingale, and thus have an increasing trend. The previsible trend part A_t of Y_t^2 is called the *quadratic variation*, for the following reason:

$$\Delta A_t = \mathbb{E}[Y_t^2 - Y_{t-1}^2 | \mathcal{F}_{t-1}] = \mathbb{E}[(Y_t - Y_{t-1})^2 | \mathcal{F}_{t-1}] = \mathbb{E}[(\Delta Y_t)^2 | \mathcal{F}_{t-1}]$$

so that $A_t = \sum_{s=1}^t \mathbb{E}[(\Delta Y_s)^2 | \mathcal{F}_{s-1}]$. Prove this.

◻

In the continuous-time theory, the generalization of the Doob decomposition to the Doob–Meyer Decomposition Theorem for submartingales is a deep result. The quadratic variation process associated with a submartingale is of great importance in deriving a general theory of stochastic integration.

Definition 11.2.11 If C is a previsible process, and if X is adapted (both with respect to a filtration $\mathcal{F}_n, n \in \mathbb{N}$), then the **martingale transform** of X by C is the process W given by

$$W_0 = 0$$

$$W_n = \sum_{k=1}^n C_k(X_k - X_{k-1}) \quad \text{if } n > 0$$

The process W is generally denoted by $C \cdot X$, and W_n by $(C \cdot X)_n$.

Thus the martingale transform of X by C is simply your winnings process on the game X using the gambling strategy C . Now comes the crunch:

Theorem 11.2.12 (a) Suppose that X is a martingale, and that C is a bounded previsible process. Then $C \cdot X$ is a martingale.

(b) If X is a supermartingale (submartingale), and C is a bounded non-negative previsible process, then $C \cdot X$ is a supermartingale (submartingale).

Proof: (a) Let $W = C \cdot X$. The fact that C is bounded and that each X_n is integrable implies that W_n is integrable as well. Then $W_{n+1} - W_n = C_{n+1}(X_{n+1} - X_n)$. Using the fact that X_n and C_{n+1} are \mathcal{F}_n -measurable, we see that

$$\mathbb{E}[W_{n+1} - W_n | \mathcal{F}_n] = C_{n+1}[\mathbb{E}[X_{n+1} | \mathcal{F}_n] - X_n] = 0$$

so that $\mathbb{E}[W_{n+1} | \mathcal{F}_n] = \mathbb{E}[W_n | \mathcal{F}_n] = W_n$.

The proof of (b) is left as an exercise.

–

This theorem has the following important consequence for games of chance: You cannot find a previsible trading strategy which will turn a fair game to your advantage, i.e. which will turn a martingale into a submartingale. No matter what your strategy, your winnings process will still be a martingale.

As a final remark, note that if X is a (super-, sub-)martingale, and a is a constant, then $Y = X + a$ is a (super-, sub-)martingale, and moreover $C \cdot X = C \cdot Y$.

Chapter 12

PDEs in Finance, with a Detour Through Black–Scholes

In this chapter we derive the Black–Scholes equations for European options. The Black–Scholes price is not model independent, i.e. it depends on the model we chose for stock prices. Accordingly, the first section of this chapter is concerned with developing a model of stock price behaviour. In the second section, we develop the machinery of stochastic calculus in an intuitive and non-rigorous manner. It should be pointed out that the motivation for Ito’s formula provided in this section is pretty flimsy, and we do not claim to give a mathematically accurate account.

Having Itô’s formula at our disposal, we then derive the Black–Scholes PDE, again in an intuitive non-rigorous manner. Instead of solving the PDE directly — we will do that later — we note that the PDE has a surprising property: The drift of the underlying asset does not occur in the PDE. This allows us to use the machinery of *risk-neutral valuation* to derive the Black–Scholes option prices.

12.1 Modelling Stock Prices

Any model of stock price behaviour must be *stochastic*, i.e. it must incorporate the random nature of price behaviour. The simplest such models are *random walks*: Let $X_t, t = 1, 2, \dots$ be a family of distributed random variables, and let S_0 be the stock price at $t = 0$. We might (naively) attempt to model the stock price process by

$$S_t = S_{t-1} + X_t \quad \text{i.e.} \quad S_t = S_0 + \sum_{u=1}^t X_u$$

The intuition behind this is that the price at time t equals the price at time $t - 1$ plus a “random shock”, modelled by X_t .

We should also assume that these shocks are *independent*. Why? If we could predict today that the stock price is going to go up *tomorrow*, this makes the stock more attractive today. Thus more people would buy it today, forcing the stock price up *today*, until it reaches

the level predicted. Thus any change in the stock price must essentially be unpredictable. This is just a version of *Efficient Markets Hypothesis*, which, loosely, asserts that all available information about a corporation is instantly reflected in its stock price. Thus future changes in price are not dependent on past changes in price.

There are several reasons why a random walk model of stock prices is inadequate, but an obvious one is that it doesn't take into account scale. For stock prices, we expect the change in price to be *proportional* to the current price. To see this, consider two companies in two parallel universes, A and B. The universes and the companies are identical, except for one thing. In universe A, the company has issued 100 shares, each trading at \$100. In universe B, the company has undertaken a 2-for-1 stock split, so that it has issued 200 shares, each trading at \$50. Both companies are otherwise identical, e.g. they are both worth \$10 000. One day an earthquake causes massive damage, and both companies lose half their value. The shares in universe A now trade at \$50, whereas those in universe B trade at \$25. Thus the share price has not dropped by the same amount in both universes: Each share has lost the same *proportion* of its value.

Simply put, if investors require a return of 14%, then they require that return irrespective of whether the share price is \$50 or \$100.

The shares of A, B change by the same *factor*, i.e. they have exactly the same change in returns (but not the same absolute change in price). This is reflected in, e.g., the binomial model, where shares can go up by a *factor* of u or down by a *factor* of $\frac{1}{u}$. But a multiplicative change in the stock price amounts to an additive change in the logarithm of the stock price:

$$S_{t+\Delta t} = u^{\pm 1} S_t \quad \text{implies} \quad \ln S_{t+\Delta t} = \ln S_t \pm \ln u$$

i.e. if we define the returns process R_t by $S_t = S_0 e^{R_t}$ (i.e. $R_t := \ln \frac{S_t}{S_0}$), and define $\delta := \ln u$, we have i.e.

$$R_{t+\Delta t} = R_t \pm \delta$$

A better random model of stock prices is therefore one in which the *returns* process R_t follows a random walk.

12.1.1 Modelling Returns in Continuous-Time

We now seek a continuous-time version of the random walk — a stochastic process that is changing because of random shocks at every instant in time. Consider a time interval $[0, T]$ and let N be a (large) integer. Define $\Delta t := \frac{T}{N}$. Let $X_n, n = 1, 2, 3, \dots$ be independent Bernoulli random variables with

$$\mathbb{P}(X_n = \Delta x) = p \quad \text{and} \quad \mathbb{P}(X_n = -\Delta x) = 1 - p =: q$$

where $\Delta x > 0$. For $t = 0, \Delta t, 2\Delta t, \dots, N\Delta t = T$, let $R_t := \sum_{i=1}^n X_n$, where $t = n\Delta t$. Thus R_t is a random walk, and

$$R_{t+\Delta t} = R_t \pm \Delta x$$

Some simple calculations yield

$$\mathbb{E}[R_t] = n(p - q)\Delta x = (p - q)\frac{\Delta x}{\Delta t}t \quad \text{Var}(R_t) = n(\Delta x^2 - (p - q)^2\Delta x^2) = 4pq\frac{\Delta x^2}{\Delta t}t$$

Now suppose we can observe the process R_t and want $\mathbb{E}[R_t] = \mu t$ and $\text{Var}(R_t) = \sigma^2 t$, where μ, σ are constants, and $\sigma > 0$. (We want $\sigma^2 > 0$, otherwise $\text{Var}(R_t) = 0$, in which case R_t would be non-random.)

In the continuous limit, i.e. as $N \rightarrow \infty$ and $\Delta t \rightarrow 0$, we must have

$$(p - q) \frac{\Delta x}{\Delta t} \rightarrow \mu \quad 4pq \frac{\Delta x^2}{\Delta t} \rightarrow \sigma^2$$

The first equation yield $\Delta x \approx \frac{\mu \Delta t}{p - q}$ when Δt is small. Substituting into the second equation, we see that

$$\frac{4pq}{(p - q)^2} \Delta t \approx \frac{\sigma^2}{\mu^2}$$

when Δt is small. Now since, $\Delta t \rightarrow 0$, we must have $\frac{4pq}{(p - q)^2} \rightarrow \infty$, for otherwise the product $\frac{4pq}{(p - q)^2} \Delta t$ would tend to 0, not $\frac{\sigma^2}{\mu^2}$. It is therefore necessary that $p - q \rightarrow 0$, and thus p, q must both tend to $\frac{1}{2}$ as $\Delta t \rightarrow 0$. From the fact that $4pq \frac{\Delta x^2}{\Delta t} \rightarrow \sigma^2$, we then see that we must have

$$\Delta x \approx \sigma \sqrt{\Delta t}$$

for small Δt .

We had $\Delta x \approx \frac{\mu \Delta t}{p - q}$ for small Δt , and thus $p - q \approx \frac{\mu}{\sigma} \sqrt{\Delta t}$. Since $p + q = 1$, we must have

$$p = \frac{1}{2} \left(1 + \frac{\mu}{\sigma} \sqrt{\Delta t} \right) \quad = \frac{1}{2} \left(1 - \frac{\mu}{\sigma} \sqrt{\Delta t} \right)$$

As a check, note that

$$\mathbb{E}[R_t] = (p - q) \frac{\Delta x}{\Delta t} t = \frac{\mu}{\sigma} \sqrt{\Delta t} \frac{\sigma \sqrt{\Delta t}}{\Delta t} t = \mu t$$

and

$$\text{Var}(R_t) = 4pq \frac{\Delta x^2}{\Delta t} t = (1 - \frac{\mu^2}{\sigma^2} \Delta t) \frac{\sigma^2 \Delta t}{\Delta t} t = \sigma^2 t - \mu^2 t \Delta t \rightarrow \sigma^2 t$$

as should be the case.

We now have an idea of how to create a continuous-time stochastic process R_t as the $(\Delta t \rightarrow 0)$ -limit of a random walk. But the limit process has some peculiar features. For example

$$\Delta R_t \approx \pm \sigma \sqrt{\Delta t} \quad \text{is of the order of } \sqrt{\Delta t}$$

If $f(t)$ is a differentiable function, then

$$\Delta f(t) \approx f'(t) \Delta t \quad \text{is of the order of } \Delta t$$

. Now when Δt is small, we see that $\sqrt{\Delta t}$ is much larger than Δt (Take, e.g. $\Delta t = 10^{-2n}$ and note that $\sqrt{\Delta t} = 10^{-n} = 10^n \Delta t$.) It follows that R_t cannot be differentiable as a function of t .

The probabilist will immediately want to know the distribution of R_t . Let $u(t, x)$ be the density of the random variable R_t , i.e.

$$u(t, x) \Delta x \approx \mathbb{P}(R_t \in [x, x + \Delta x])$$

At time $t + \Delta t$ the random walk can reach the point x in two ways: It can move right from the point $x - \Delta x$ at time t , with probability p , or it can move left from the point $x + \Delta x$, with probability q . Thus

$$u(t + \Delta t, x) = pu(t, x - \Delta x) + qu(t, x + \Delta x)$$

Now we Taylor expand up to order Δt . Firstly

$$u(t + \Delta t, x) \approx u(t, x) + u_t(t, x)\Delta t + o(\Delta t)$$

Next,

$$u(t, x \pm \Delta x) = u(t, x) \pm u_x(t, x)\Delta x + \frac{1}{2}u_{xx}(t, x)\Delta x^2 + o(\Delta x^2)$$

Here, we have taken a second-order Taylor expansion, because Δx is of the order $\sqrt{\Delta t}$, and Δx^2 of the order Δt . Putting these together, we obtain (at the point (t, x)):

$$u + u_t\Delta t = (p + q)u + (-p + q)u_x\Delta x + \frac{1}{2}(p + q)u_{xx}\Delta x^2$$

However, we know that $p = \frac{1}{2}(1 + \frac{\mu}{\sigma}\sqrt{\Delta t})$ and that $\Delta x \approx \sigma\sqrt{\Delta t}$ and $p, q \rightarrow \frac{1}{2}$. Hence

$$u_t\Delta t = -(\frac{\mu}{\sigma}\sqrt{\Delta t})u_x(\sigma\sqrt{\Delta t}) + \frac{1}{2}u_{xx}\sigma^2\Delta t$$

which yields the following partial differential equation for the density of R_t .

$$u_t = -\mu u_x + \frac{1}{2}\sigma^2 u_{xx}$$

However, the PDE is not sufficient to determine the density u : It has many solutions. We seek a solution which has the following properties:

- For each $t \geq 0$, we have $\int_{-\infty}^{\infty} u(t, x) dx = 1$, because $u(t, x)$ is a density, and
- $u(0, x)$ is rather odd: We have $R_0 = 0$, and so

$$f(0) = \mathbb{E}[f(R_0)] = \int_{-\infty}^{\infty} f(x)u(0, x) dx$$

i.e. $u(0, x)$ is a “function” with the property that $\int_{-\infty}^{\infty} f(x) dx = f(0)$ for every function f . The “function” with this property is called the *Dirac delta* δ_0 . It is not a function at all (but the simplest example of a so-called *generalized function* or *distribution* (in the sense of Schwartz).) Nevertheless, we can get some intuition as to how u ought to behave. We see that for t close to 0, the density $u(t, x)$ must be very small for $x \neq 0$, because R_t must be close to x when t is near zero. Yet the area under the curve is 1, i.e. $u(t, x)$ must be extremely peaked at around $x = 0$ and then rapidly drop off. We may thus think off $u(0, x) = \delta_0$ as a “function” which has

$$\delta_0(x) = 0 \text{ when } x \neq 0 \quad \delta_0(0) = +\infty \text{ in such a way that } \int_{-\infty}^{\infty} \delta_0(x) dx = 1$$

Oddly enough, we can find such a function. The PDE for the density, derived by Einstein in 1905, is a version of the heat equation, derived by Fourier, which governs heat transfer. So this PDE was not new: It had been intensively studied by physicists, with $u(t, x)$ playing

the role of the temperature at time t at a point x in an infinitely long rod. The *fundamental solution* or *Green's function* of such a PDE was well-known

$$u(t, x) = \frac{1}{\sqrt{2\pi\sigma^2 t}} e^{-\frac{(x-\mu t)^2}{2\sigma^2 t}}$$

We will give a derivation of this result later on, but you can verify by direct differentiation that this function does, in fact, satisfy the PDE. You will also immediately recognize it as the density of an $N(\mu t, \sigma^2 t)$ -random variable. Furthermore, for t near 0, such a random variable has very small standard deviation, and thus the density is extremely peaked around 0, just as we require.

It follows, therefore, that the density of t is $N(\mu t, \sigma^2 t)$. Of course, the Central Limit Theorem states that, subject to a moment condition, large sums of i.i.d. are roughly normally distributed, so we are not surprised. But here, we have in essence given a proof of the Central Limit Theorem by PDE methods, at least for random walks of the type described.

When we take $\mu = 0$ and $\sigma = 1$, we obtain one of the basic building blocks of financial modelling:

Definition 12.1.1 Standard *Brownian motion* is a continuous-time stochastic process $B_t, t \geq 0$ with the following properties:

(1) Each change

$$B_t - B_s = (B_{s+h} - B_s) + (B_{s+2h} - B_{s+h}) \\ + \cdots + (B_t - B_{t-h})$$

is *normally distributed* with mean 0 and variance $t - s$.

(2) Each change $B_t - B_s$ is **independent** of all the previous values $B_u, u \leq s$.

(3) Each sample path $B_t, t \geq 0$ is (a.s.) **continuous**, and has $B_0 = 0$.

Now put

$$R_t = \mu t + \sigma B_t$$

It then follows easily that

$$R_t \sim N(\mu t, \sigma^2 t)$$

i.e. the standard Brownian motion can also be used to model returns processes where $\mu \neq 0$ and $\sigma \neq 1$. The process R_t is called an *arithmetic Brownian motion* with drift rate μ and variance rate σ^2 . We will also refer to σ as the *volatility*.

12.1.2 Modelling Share Prices in Continuous Time

We have obtained a model for share prices:

$$S_t = e^{R_t} = e^{\mu t + \sigma B_t}$$

We shall soon see that this translates to a *stochastic differential equation*

$$dS_t = \alpha S_t dt + \sigma S_t dB_t \quad \alpha := \mu + \frac{1}{2}\sigma^2$$

i.e. the proportional change in share price $\frac{dS_t}{S_t}$ can be decomposed into two terms, αdt and σdB_t . Such a process is called a *geometric Brownian motion* (GBM) with *drift* α and

volatility σ . The drift is the (proportional) rate at which the share price increases in the absence of risk. The differential dB_t models the randomness (risk), and the volatility models how sensitive the share price is to these random events. The greater α , the faster the share price increases in the absence of risk. The greater σ , the more violently the share price reacts to random events. Note that dB_t can be negative (unlike dt), allowing for decreases in share price. Also note that $|dB_t| \approx \sqrt{dt} \gg dt$, so that over short periods the change in share price is dominated by random events. Many of these random events cancel out however, so that in the long run the drift term is dominant.

Now consider a market with a share S_t whose price follows a GBM $dS_t = \alpha S dt + \sigma S dB_t$. Let the risk-free interest rate be r , i.e. the risk-free bank account A_t satisfies the DE

$$dA_t = rA_t dt$$

A_t is the *riskless asset*. It has drift r and zero volatility.

A portfolio is a two-dimensional process (θ_t^0, θ_t^1) , where θ_t^1 is the number of shares owned at time t , θ_t^0 is the amount of money in the bank account at time t , discounted to time 0. Given such a dynamic portfolio $\theta_t = (\theta_t^0, \theta_t^1)$, the *value process* $V_t(\theta)$ satisfies

$$\begin{aligned} dV_t &= \theta_t^0 dA_t + \theta_t^1 dS_t \\ &= (r\theta_t^0 A_t + \mu\theta_t^1 S_t) dt + \theta_t^1 \sigma S_t dB_t \end{aligned}$$

The value of the portfolio at time T is therefore

$$\begin{aligned} V_T(\theta) &= V_0(\theta) + \int_0^T [r\theta_t^0 A_t + \mu\theta_t^1 S_t] dt \\ &\quad + \int_0^T \theta_t^1 \sigma S_t dB_t \end{aligned}$$

We now see that we need to be able to evaluate integrals of the form

$$\int_0^T f(t) dB_t$$

This is an example of a *stochastic integral*. The obvious method would be to regard the above as a Riemann–Stieltjes (or Lebesgue–Stieltjes) integral. However, it can be shown that this approach will not work. Nevertheless, it is possible to define the stochastic integrals, and there is even a very simple rule which allows us to manipulate them: Itô's formula. However, the rules of stochastic calculus do differ from those of ordinary calculus. We are, after all, now working with stochastic processes whose paths are nowhere differentiable, whereas ordinary calculus deals only with differentiable functions.

12.2 A Naive Approach to Stochastic Calculus

Let $f(x)$ be a differentiable function on an interval $[a, b]$. Partition this interval:

$$a = x_0 < x_1 < x_2 < \dots < x_n = b$$

where $x_{i+1} - x_i = \Delta x$. Then by Taylor series expansion, we get

$$\begin{aligned} f(x_{i+1}) - f(x_i) &= f'(x_i)\Delta x + \frac{1}{2!}f''(x_i)(\Delta x)^2 \\ &\quad + \frac{1}{3!}f'''(x_i)(\Delta x)^3 + \text{terms involving } \Delta x^4, \Delta x^5, \dots \end{aligned}$$

Thus

$$\begin{aligned} f(b) - f(a) &= \sum_{i=0}^{n-1} [f(x_{i+1}) - f(x_i)] \\ &= \sum_{i=0}^{n-1} f'(x_i) \Delta x + \frac{1}{2} \sum_{i=0}^{n-1} f''(x_i) (\Delta x)^2 + \dots \end{aligned}$$

As $\Delta x \rightarrow 0$, we get

$$\begin{aligned} f(b) - f(a) &= \lim_{\Delta x \rightarrow 0} \sum_i f'(x_i) \Delta x + \frac{1}{2} \lim_{\Delta x \rightarrow 0} \sum_i f''(x_i) (\Delta x)^2 \\ &\quad + \dots \\ &= \int_a^b f'(x) dx + \left[\frac{1}{2} \int_a^b f''(x) (dx)^2 + \dots \right] \end{aligned}$$

In ordinary calculus, only the first term counts (by the Fundamental Theorem of Calculus), and the other terms are zero. This is because the *quadratic variation* of any “ordinary” function is zero, i.e.

$$\lim_{\Delta x \rightarrow 0} \sum (\Delta g)^2 = 0$$

for any “ordinary” function g . This is not all that hard to see: We have

$$\begin{aligned} \lim_{\Delta x \rightarrow 0} \sum (\Delta g)^2 &= \lim_{\Delta x \rightarrow 0} \sum g'(x)^2 \Delta x^2 \\ &= \left(\lim_{\Delta x \rightarrow 0} \Delta x \right) \left(\lim_{\Delta x \rightarrow 0} \sum g'(x)^2 \Delta x \right) \\ &= 0 \cdot \int_a^b g'(x)^2 dx \\ &= 0 \end{aligned}$$

(assuming that g is continuously differentiable).

But Brownian motion is different: Consider $\Delta B = B_{t+\Delta t} - B_t$. This is a normally distributed random variable with $\mathbb{E}[\Delta B] = 0$ and variance $\text{var}(\Delta B) = \Delta t$.

Consider next the random variable $(\Delta B)^2$. This has

$$\begin{aligned} \mathbb{E}[(\Delta B)^2] &= \text{var}[\Delta B] = \Delta t \\ \text{var}[(\Delta B)^2] &= \mathbb{E}[(\Delta B)^4] - (\Delta t)^2 = 2(\Delta t)^2 \ll \Delta t \end{aligned}$$

Thus the variance of $(\Delta B)^2$ is ≈ 0 , i.e. though ΔB is a random variable, $(\Delta B)^2$ is a constant.¹ It follows that

$$\lim_{\Delta t \rightarrow 0} \sum \mathbb{E}(\Delta B)^2 = \lim_{\Delta t \rightarrow 0} \sum \Delta t = T$$

where T is the total elapsed time. Thus the quadratic variation of Brownian motion is non-zero.

Also

$$\lim_{\Delta t \rightarrow 0} \sum \mathbb{E}(\Delta B)^4 = 2 \lim_{\Delta t \rightarrow 0} \sum (\Delta t)^2 = 0$$

¹This nonsense can be made precise, promise.

because $g(t) = t$ is an “ordinary” function, with quadratic variation zero. Hence we cannot ignore the second-order term

$$\frac{1}{2} \int_a^b f''(x) (dx)^2$$

in the case that $x = B$, but we can ignore all higher-order terms.

We thus have the following rules for stochastic calculus:

$$\begin{aligned} (dB_t)^2 &= dt \\ dB_t \cdot dt &= (dt)^2 = 0 \end{aligned}$$

Suppose that $f(t, x)$ is a $C^{1,2}$ -function, and let $X_t = f(t, B_t)$. Applying these rules to a second order Taylor series, we obtain:

Theorem: (Ito’s Formula)

$$dX_t = \left(\frac{\partial f}{\partial t} + \frac{1}{2} \frac{\partial^2 f}{\partial B^2} \right) dt + \frac{\partial f}{\partial B} dB_t$$

Ordinary calculus shows that for a function $f(t, x)$ we have

$$df = \frac{\partial f}{\partial t} dt + \frac{\partial f}{\partial x} dx$$

In stochastic calculus, we get another term, due to the non-zero quadratic variation of Brownian motion.

Example 12.2.1 Take our model for stock prices in terms of the returns process:

$$S_t = e^{R_t} \quad R_t = \mu t + \sigma B_t$$

Using Itô’s formula, we get

$$dS_t = 0 dt + e^{R_t} dR_t + \frac{1}{2} e^{R_t} (dR_t)^2 = S_t [dR_t + \frac{1}{2} (dR_t)^2]$$

Now

$$dR_t = \mu dt + \sigma dB_t \quad (dR_t)^2 = \sigma^2 dt$$

so

$$dS_t = S_t [\mu dt + \sigma dB_t + \frac{1}{2} \sigma^2 dt] = (\mu + \frac{1}{2} \sigma^2) S_t dt + \sigma S_t dB_t$$

as claimed earlier.

□

Now let’s have another look at volatility. The GBM model for stock prices is

$$dS_t = \alpha S_t dt + \sigma S_t dB_t$$

Thus

$$\mathbb{E} \left[\frac{dS}{S} \right]^2 = \sigma^2 dt$$

and thus $\sigma^2 dt$ is the variance of the return of the stock over a small period dt .

It follows that σ is the standard deviation of the annual return of the stock S . This can be measured from market data.

Can we also measure the drift α ? This is much harder², because over short periods, the dB_t -term dominates the dt -term. The “correct” real-world dynamics of a share price is *difficult to estimate*: We can get the volatility, but not the drift. Amazingly, *we don't care*, as you will see shortly. To calculate option values we need only the volatility, not the drift.

12.3 The Black–Scholes Model

12.3.1 The Black-Scholes PDE

Using Ito's formula, it is not hard to derive a partial differential equation for European style derivatives.

Consider again market with a share S_t whose price process satisfies the SDE

$$dS = \mu S dt + \sigma S dB_t$$

Let the risk-free interest rate be r , and let A_t be the riskless bank account, with dynamics

$$dA_t = rA_t dt$$

Let $V(t, S_t)$ be European-style derivative whose value depends on both the share price and time. Consider a portfolio Π which contains 1 derivative, and n shares, i.e. its value is

$$\Pi_t = V_t + nS_t$$

A small amount of time dt later, the share price has changed. The value of the portfolio changes by

$$d\Pi_t = dV_t + n dS_t$$

By Ito's Formula,

$$\begin{aligned} dV_t &= \frac{\partial V}{\partial t} dt + \frac{\partial V}{\partial S} dS + \frac{1}{2} \frac{\partial^2 V}{\partial S^2} dS^2 \\ &= \left(\frac{\partial V}{\partial t} + \mu S \frac{\partial V}{\partial S} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} \right) dt + \sigma S \frac{\partial V}{\partial S} dB_t \end{aligned}$$

Hence

$$\begin{aligned} d\Pi_t &= \left(\frac{\partial V}{\partial t} + \mu S \frac{\partial V}{\partial S} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + n\mu S \right) dt \\ &\quad + \sigma S \left(\frac{\partial V}{\partial S} + n \right) dB_t \end{aligned}$$

Thus

$$\begin{aligned} d\Pi_t &= \left(\frac{\partial V}{\partial t} + \mu S \left[\frac{\partial V}{\partial S} + n \right] + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} \right) dt \\ &\quad + \sigma S \left[\frac{\partial V}{\partial S} + n \right] dB_t \end{aligned}$$

²Martin Davis once claimed, in a talk that I attended, that one would need 1500 years of data to get a reasonably accurate estimate of the drift — I'm no statistician, though.

Now if we take $n = -\frac{\partial V}{\partial S}$ (i.e. the portfolio is short $-\frac{\partial V}{\partial S}$ shares), then the portfolio is unaffected by the random changes in stock prices:

$$d\Pi_t = \left(\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} \right) dt \quad (12.1)$$

Thus, for a brief moment, the portfolio is risk-free. By a *no-arbitrage argument*, it must *earn the same return as the risk-free bank account*³, i.e.

$$d\Pi_t = r\Pi_t dt = r \left(V - \frac{\partial V}{\partial S} S \right) dt \quad (12.2)$$

Equating (12.1) and (12.2), we get

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS \frac{\partial V}{\partial S} - rV = 0$$

This is the famous **Black-Scholes** PDE. It is a second-order parabolic PDE, i.e. essentially a heat equation. Most of the PDE's encountered in finance are of a similar type.

Note that if a portfolio contains $\frac{\partial V}{\partial S}$ shares, then the change in the portfolio value is the same as the change in the value of the derivative. The quantity $\frac{\partial V}{\partial S}$ is called the *delta* of the derivative. One can thus *synthetically replicate* any European style derivative with underlying share S by holding, at any time, delta-many shares. This procedure is called *delta hedging*.

Consider a European call option C on a share S with strike K and maturity T . The volatility of the underlying share S is σ and the risk-free rate is r . To find the value of the call option, we must solve the following boundary value problem:

$$\begin{cases} \frac{\partial C}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 C}{\partial S^2} + rS \frac{\partial C}{\partial S} - rC = 0 \\ C(T) = \max\{S_T - K, 0\} \end{cases}$$

It is now clear why we don't care about the drift μ of the underlying asset S : It does not appear in the Black-Scholes PDE, and is therefore irrelevant to pricing derivatives.

12.3.2 Pricing in the Risk-Neutral World

In this section we calculate the Black-Scholes prices of vanilla European options. However, we use a slightly subtle probabilistic argument, rather than a brute force "solve the PDE" approach.

In the previous section, we deduced the Black-Scholes PDE for a European-style derivative V :

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS \frac{\partial V}{\partial S} - rV = 0$$

Note once again that the *drift* μ *does not occur* in the Black-Scholes PDE, though the volatility σ *does* appear. Hence the price of V is independent of μ , i.e. different values of μ will give the same price.

Thus, for example, the price of a call option on a share S with a given strike K and maturity T will be the same whether μ is very small or very big. This may seem *counterintuitive*,

³This is the *cruz* of the argument!

because if μ is very big, it seems as though the option is more likely to expire in-the-money. One would therefore think that the call option price should be an increasing function of μ . But your intuition is just plain wrong.

Since we don't care about the drift rate μ of an underlying asset, we may as well simplify our asset price dynamics by assuming that *all assets have the same drift*. Now the riskless asset (bank account) has drift r , and r occurs in the Black–Scholes PDE. We can't change the drift of the riskfree bank account without changing the PDE, and thus the solution to the pricing problem. So if we want to assume that all assets have the same drift, we have to assume that the drift of all assets is the risk-free rate r .

Mathematically, this corresponds to a change of measure — from a real world, unknowable probability measure \mathbb{P} to a knowable, *risk-neutral measure* \mathbb{Q} . In the risk-neutral world, the dynamics of S are

$$dS_t = rS_t dt + \sigma S dB_t$$

Thus we change the drift of the asset from μ to r .

Why is the world where all assets have the same return called the *risk-neutral* world? Ordinarily, investors are influenced by risk: They weigh up the expected return against the riskiness of an investment. Generally, investors are *risk averse*, which means that they require a premium in order to take on risk. Thus assets with greater riskiness (= volatility) have a higher (“real world”) expected payoff than assets with less risk. In the risk-neutral world, investors are indifferent to risk, i.e. they do not require a risk premium. The only thing they care about is *expected return*. Such investors will buy assets with a higher expected return, and sell assets with a lower expected return, regardless of the risk involved. Prices will thus adjust so that all assets have the *same* expected return (in equilibrium). Thus in a world where all investors are risk-neutral, all assets will have the same expected return, i.e. the same expected return as the risk-free bank account.

To summarize, prices in the real- and risk-neutral world are the same. It is just probabilities that are changed. Now we can calculate option prices in the risk-neutral world, because the asset price dynamics are known, and so is the distribution of future stock prices.

Now suppose that we can find a portfolio Π of traded assets which exactly hedges the payoff of a European style derivative V , so that

$$\Pi_T = V_T$$

at the derivative's maturity T . Such a portfolio is called a *replicating portfolio*. By the Law of One Price, therefore, we must have $\Pi_0 = V_0$, where Π_0, V_0 are, respectively the values of the replicating portfolio and the derivative at $t = 0$. Thus:

*If a derivative has a replicating portfolio, then
the value of the derivative equals the value of
the replicating portfolio.*

Now in the Black–Scholes model, any European style derivative has a replicating portfolio: A portfolio consisting, at any time, of $\Delta = \frac{\partial V}{\partial S}$ shares will exactly replicate the derivative V (delta hedging).

Π_T and V_T are random variables. But since they are identical, they must have the same expectation, in any world. Since the expected return of all traded assets is r in the risk-neutral world, and since Π consists entirely of traded assets, the expected return of Π is also

r :

$$\mathbb{E}_{RN}[\Pi_T] = \Pi_0 e^{rT}$$

where Π_0 is the value of the portfolio at $t = 0$. Now since $\Pi_0 = V_0$ (by the Law of One Price) and $\Pi_T = V_T$ (because Π is a replicating portfolio of V), we see that

$$V_0 = e^{-rT} \mathbb{E}_{RN}[V_T]$$

We have therefore discovered the following procedure for valuing a derivative V .

- (i) Assume that the drift of the underlying asset S is r (instead of μ). This moves us from the real world to the risk-neutral world.
- (ii) Calculate the expected pay-off of V at maturity in the *risk-neutral world*: $\mathbb{E}_{RN}[V_T]$
- (iii) Discount to the present to get the price today:

$$V_0 = e^{-rT} \mathbb{E}_{RN}[V_T]$$

The point is that we can't calculate $\mathbb{E}_{\text{real}}[V_T]$, because we do not know the distribution of the underlying S_T in the real world. However, we *can* calculate $\mathbb{E}_{RN}[V_T]$: Since we know the drift of S_T in the risk-neutral world, we can calculate the distribution of S_T here. This brings us to our next topic.

12.3.3 The Distribution of Asset Prices

We have postulated an asset price model

$$\frac{dS_t}{S_t} = \mu dt + \sigma dB_t$$

where $\mu = r$ in the risk-neutral world. Consider now the function $Y_t = f(S_t) = \ln S_t$. By Ito's formula,

$$\begin{aligned} dY_t &= \frac{1}{S_t} dS_t - \frac{1}{2} \frac{1}{S_t^2} (dS_t)^2 \\ &= \left(\mu - \frac{1}{2}\sigma^2\right) dt + \sigma dB_t \end{aligned}$$

using $(dB_t)^2 = dt$, $(dt)^2 = 0 = (dB_t)(dt)$.

So $Y_t = \ln S_t$ follows a *Brownian motion with drift*. We can easily solve this SDE to get

$$Y_T - Y_0 = \left(\mu - \frac{1}{2}\sigma^2\right)T + \sigma B_T$$

which implies that Y_T is normally distributed with mean $Y_0 + (\mu - \frac{1}{2}\sigma^2)T$ and variance $\sigma^2 T$:

$$Y_T \sim N\left(Y_0 + \left(\mu - \frac{1}{2}\sigma^2\right)T, \sigma^2 T\right)$$

Thus the log of the stock price is normally distributed. We say that stock prices are *lognormally distributed* (in the Black-Scholes model).

Definition 12.3.1 A random variable X is said to be lognormally distributed if and only if the random variable $Y = \ln X$ is normally distributed. Equivalently, if X is normally distributed, then e^X is lognormally distributed.

□

The density function of a lognormal variable

Suppose that $X \sim N(\mu_X, \sigma_X^2)$, so that X has density

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma_X^2}} e^{-(x-\mu_X)^2/2\sigma_X^2}$$

Let $Y = e^X$, so that Y is lognormally distributed. Let F_Y, f_Y be, respectively, the distribution and density functions of Y . Then

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X \leq \ln y) = F_X(\ln y)$$

Differentiating,

$$f_Y(y) = F'_Y(y) = F'_X(\ln y) \frac{1}{y} = \frac{1}{y} f_X(\ln y)$$

Thus:

The density of a lognormal random variable Y is given by

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma_X^2}} \frac{1}{y} \exp \left[-\frac{(\ln y - \mu_X)^2}{2\sigma_X^2} \right] & \text{if } y > 0 \\ 0 & \text{if } y \leq 0 \end{cases}$$

where $Y = \ln X$ and $X \sim N(\mu_X, \sigma_X^2)$. Moreover, the mean μ_Y and variance σ_Y^2 of Y are given by

$$\mu_Y = e^{\mu_X + \frac{1}{2}\sigma_X^2} \quad \sigma_Y^2 = e^{2\mu_X + \sigma_X^2} [e^{\sigma_X^2} - 1]$$

The statements about the mean and variance of a lognormal random variables are left as straightforward exercises in integration. For example, you must show that

$$\mathbb{E}[Y] = \frac{1}{\sqrt{2\pi\sigma_x^2}} \int_0^\infty \frac{1}{y} e^{-\frac{(\ln y - \mu_x)^2}{2\sigma_x^2}} dy = e^{\mu_x + \frac{1}{2}\sigma_x^2}$$

We can now prove the following important fact:

Theorem 12.3.2 Suppose that Y is lognormally distributed, where $\ln Y \sim N(m, s^2)$. Let K be a positive constant. Then

$$\begin{aligned} \mathbb{P}(Y \geq K) &= N(d_-) \\ \mathbb{E}[\max\{Y - K, 0\}] &= \mathbb{E}[Y]N(d_+) - KN(d_-) \end{aligned}$$

where

$$d_{\pm} = \frac{\ln[\mathbb{E}(Y)/K] \pm \frac{1}{2}s^2}{s}$$

Proof: Since $\ln Y \sim N(m, s^2)$, it follows that if we define $X = \frac{\ln Y - m}{s}$, then $X \sim N(0, 1)$, i.e. X is a standard normal random variable. Clearly

$$\begin{aligned}\mathbb{P}(Y \geq K) &= \mathbb{P}(\ln Y \geq \ln K) \\ &= \mathbb{P}\left(X \geq \frac{\ln K - m}{s}\right) \\ &= 1 - N\left(\frac{\ln K - m}{s}\right) \\ &= N\left(\frac{m - \ln K}{s}\right)\end{aligned}$$

where $N(x)$ is the distribution function of a standard normal random variable, and we used the fact that $1 - N(x) = N(-x)$.

But we know that $\mathbb{E}[Y] = e^{m + \frac{1}{2}s^2}$, so that $m = \ln \mathbb{E}[Y] - \frac{1}{2}s^2$. We thus obtain

$$\mathbb{P}(Y \geq K) = N\left(\frac{\ln \mathbb{E}[Y] - \ln K - \frac{1}{2}s^2}{s}\right) = N(d_-)$$

as required.

Now $\mathbb{E}[\max\{Y - K, 0\}]$ is an integral which can be split up into two parts. In the first region, $Y \geq K$, so that $\max\{Y - K, 0\} = Y - K$ (in that region). In the second region, $Y < K$, so that $\max\{Y - K, 0\} = 0$. Thus

$$\mathbb{E}[\max\{Y - K, 0\}] = \int_K^\infty (y - K)f(y) dy$$

where $f(y)$ is the density function of Y . It is simpler to work with X , however, so we change variables: Put $x = \frac{\ln y - m}{s}$. Then $y = e^{sx+m}$, and

$$\mathbb{E}[\max\{Y - K, 0\}] = \mathbb{E}[\max\{e^{sx+m} - K, 0\}] = \int_{(\ln K - m)/s}^\infty (e^{sx+m} - K)g(x) dx$$

where $g(x)$ is the density of the standard normal random variable X . We can split this up into two integrals:

$$(1) \int_{(\ln K - m)/s}^\infty e^{sx+m} g(x) dx$$

$$(2) -K \int_{(\ln K - m)/s}^\infty g(x) dx$$

We simplify the integrand of the first integral by completing the square:

$$\begin{aligned}e^{sx+m} g(x) &= e^{sx+m} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \\ &= \frac{1}{\sqrt{2\pi}} e^{-(x^2 - 2sx + s^2)/2} e^{m+s^2/2} \\ &= e^{m+s^2/2} g(x - s) \\ &= \mathbb{E}[Y] g(x - s)\end{aligned}$$

where we used the fact that $\mathbb{E}[Y] = e^{m+s^2/2}$. Thus the first integral becomes

$$\int_{(\ln K - m)/s}^{\infty} e^{sx+m} g(x) dx = \mathbb{E}[Y] \int_{(\ln K - m)/s}^{\infty} g(x - s) dx$$

and the $\int_a^{\infty} g(x - s) dx$ is just the probability that a standard normal random variable is greater than $a - s$, which is $1 - N(a - s) = N(s - a)$. Thus

$$\int_{(\ln K - m)/s}^{\infty} e^{sx+m} g(x) dx = \mathbb{E}[Y] N\left(s - \frac{\ln K - m}{s}\right) = \mathbb{E}[Y] N(d_+)$$

using $m = \ln \mathbb{E}[Y] - s^2/2$.

Similarly, but rather more easily, it can be shown that

$$-K \int_{(\ln K - m)/s}^{\infty} g(x) dx = -KN(d_-)$$

and this completes the proof. □

The distribution of asset prices

We have

$$\frac{dS_t}{S_t} = \mu dt + \sigma dB_t$$

which we solved to obtain

$$\ln S_t \sim N\left(\ln S_0 + \left(\mu - \frac{1}{2}\sigma^2\right)t, \sigma^2 t\right)$$

Thus $S_t = e^X$, where $X \sim N(\mu_X, \sigma_X^2)$ and

$$\begin{aligned}\mu_X &= \ln S_0 + \left(\mu - \frac{1}{2}\sigma^2\right)t \\ \sigma_X^2 &= \sigma^2 t\end{aligned}$$

So the density of S_t is

$$f(S) = \frac{1}{\sqrt{2\pi\sigma^2 t}S} e^{-\frac{[\ln S - (\ln S_0 + (\mu - \frac{1}{2}\sigma^2)t]]^2}{2\sigma^2 t}} \quad S \geq 0$$

and

$$\mathbb{E}[S_t] = e^{\ln S_0 + \mu t} = S_0 e^{\mu t}$$

Replacing μ with r will give the density of S_t in the risk-neutral world.

Example 12.3.3 Consider a (long) forward contract F on an asset with forward price $K = S_0 e^{rT}$. The payoff of F at T is $S_T - K$. Thus the value of the contract today is

$$F_0 = e^{-rT} \mathbb{E}_{RN}[S_T - K]$$

In the risk–neutral world, the asset price dynamics are given by

$$\frac{dS_t}{S_t} = r dt + \sigma dB_t \quad \text{i.e.} \quad S_t = S_0 e^{(r - \frac{1}{2}\sigma^2)t + \sigma B_t}$$

Thus $\mathbb{E}_{RN}[S_T] = S_0 e^{rT}$, and so the value of the forward contract is

$$F_0 = e^{-rT} [S_0 e^{rT} - K] = 0$$

which is the *correct* value obtained by the (presumably familiar) static replication argument.

If we had used the “real world” drift μ , however, we would have obtained

$$F_0 = e^{-rT} [S_0 e^{\mu T} - K]$$

and this is incorrect. □

Exercise 12.3.4 Recall that the value of a forward contract at time t is

$$F_t = [S_t - S_0 e^{rt}]$$

Show that F_t satisfies the BS PDE with boundary condition $F_0 = 0$. □

By the results in the previous section, we have

$$\begin{aligned} \mathbb{E} S_T &= e^{\mu T + \frac{1}{2}\sigma^2 T} \\ &= S_0 e^{\mu T} \end{aligned}$$

So μ is the expected rate of return of the asset S . In particular, in the risk–neutral world the drift of every traded asset is $\mu = r$, so in the risk–neutral world all assets have the same expected rate of return r (which we already knew).

Also, $\mathbb{E}[(\frac{dS_t}{S_t})^2] = \sigma^2 dt$, which shows that $\sigma^2 dt$ is the variance of returns over a period dt . We can thus interpret σ to be the standard deviation of the returns on S over a period of one year.

12.4 Option Pricing: The Black–Scholes Formula

Now that we have the density function of the asset price S_T in the risk–neutral world, we can price practically any European claim V with payoff $\Phi(S_T)$:

$$\begin{aligned} V_0 &= e^{-rT} \mathbb{E}_{RN}[\Phi(S_T)] \\ &= \frac{e^{-rT}}{\sqrt{2\pi\sigma^2 T}} \int_0^\infty \Phi(S) e^{-\frac{[\ln S - (\ln S_0 + (r - \frac{1}{2}\sigma^2)T)]^2}{2\sigma^2 T}} \frac{dS}{S} \end{aligned}$$

It is easy to evaluate this integral numerically, using Simpson’s method, for example.

Consider next a call option with strike K and maturity T . In this case, $\Phi(S_T) = \max\{S_T - K, 0\}$. Thus:

$$C_0 = e^{-rT} \mathbb{E}_{RN}[\max\{S_T - K, 0\}]$$

Now in the risk-neutral world, S_T is lognormally distributed, with $\ln S_T \sim N(\ln S_0 + (r - \frac{1}{2}\sigma^2)T, \sigma^2 T)$. By Theorem 12.3.2, therefore,

$$\mathbb{E}_{RN}[\max\{S_T - K, 0\}] = \mathbb{E}_{RN}[S_T]N(d_+) - KN(d_-)$$

where

$$d_{\pm} = \frac{\ln[\mathbb{E}(S_T)/K] \pm \frac{1}{2}\sigma^2 T}{\sigma\sqrt{T}}$$

But $\mathbb{E}_{RN}[S_T] = S_0 e^{rT}$, and thus (remembering to discount):

$$C_0 = S_0 N(d_+) - Ke^{-rT} N(d_-)$$

where

$$d_{\pm} = \frac{\ln \frac{S_0 e^{rT}}{K} \pm \frac{1}{2}\sigma^2 T}{\sigma\sqrt{T}}$$

and $N(x)$ is the distribution function of a standard normal random variable, i.e.

$$N(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt$$

The normal distribution function $N(x)$ can be determined from tables, or by using the Excel function NORMSDIST.

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.5279	0.53188	0.53586
0.1	0.53983	0.543795	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.659097	0.66276	0.6664	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.7054	0.70884	0.71226	0.71566	0.71904	0.7224
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.7549
0.7	0.75804	0.76115	0.76424	0.7673	0.77035	0.77337	0.77637	0.77935	0.7823	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.8665	0.86864	0.87076	0.87286	0.87493	0.87698	0.8790	0.8810	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.9032	0.9049	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.9222	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.9452	0.9463	0.94738	0.94845	0.9495	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.9608	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.9732	0.97381	0.97441	0.9750	0.97558	0.97615	0.9767
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.9803	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.9830	0.98341	0.98382	0.98422	0.98461	0.9850	0.98537	0.98574
2.2	0.9861	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.9884	0.9887	0.98899
2.3	0.98928	0.98956	0.98983	0.9901	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.9918	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.9943	0.99446	0.99461	0.99477	0.99492	0.99506	0.9952
2.6	0.99534	0.99547	0.9956	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.9972	0.99728	0.99736
2.8	0.99744	0.99752	0.9976	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.9990
3.1	0.99903	0.99906	0.9991	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.9994	0.99942	0.99944	0.99946	0.99948	0.9995
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.9996	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.9997	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.9998	0.99981	0.99981	0.99982	0.99983	0.99983
3.6	0.99984	0.99985	0.99985	0.99986	0.99986	0.99987	0.99987	0.99988	0.99988	0.99989
3.7	0.99989	0.9999	0.9999	0.9999	0.99991	0.99991	0.99992	0.99992	0.99992	0.99992
3.8	0.99993	0.99993	0.99993	0.99994	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995
3.9	0.99995	0.99995	0.99996	0.99996	0.99996	0.99996	0.99996	0.99996	0.99997	0.99997
4.0	0.99997	0.99997	0.99997	0.99997	0.99997	0.99997	0.99998	0.99998	0.99998	0.99998

Table for $N(x)$ when $x \geq 0$.**Notes:**

- (1) Note that $N(-x) = 1 - N(x)$.
- (2) Use linear interpolation to calculate the values of $N(x)$: For example,

$$N(0.8625) = N(0.86) + 0.25(0.87 - 0.86) \approx N(0.86) + 0.25[N(0.87) - N(0.86)]$$

Exercise 12.4.1 Verify that the formula for the call option above is a solution to the BS PDE. Begin by showing that

$$\begin{aligned}\Delta : \frac{\partial C}{\partial S} &= N(d_+) \\ \Gamma : \frac{\partial^2 C}{\partial S^2} &= \frac{N'(d_+)}{\sigma S \sqrt{T}} \\ \Theta : \frac{\partial C}{\partial t} &= -\frac{\sigma S N'(d_+)}{2\sqrt{T}} - rK e^{-rT} N(d_+)\end{aligned}$$

□

The partial derivatives in the above exercise are known as *the Greeks*, and are measures of the sensitivity of an option to its parameters. Other Greeks are

$$\begin{aligned}\rho : \frac{\partial C}{\partial r} \\ \text{Vega} : \frac{\partial C}{\partial \sigma}\end{aligned}$$

We can obtain the formula of a put option in the same way that we derived the formula for a call option, i.e. via a long and complicated chain of integrations. However, it is more intelligent to use *put-call parity*:

$$\begin{aligned}P_0 &= C_0 + K e^{-rT} - S_0 \\ &= S_0 [N(d_+) - 1] + K e^{-rT} [1 - N(d_-)] \\ &= -S_0 N(-d_+) + K e^{-rT} N(-d_-)\end{aligned}$$

i.e.

$$P_0 = -S_0 N(-d_+) + K e^{-rT} N(-d_-)$$

Finally, as a curiosity, we mention *binary* or *digital options*:

Definition 12.4.2 • A binary call on S with strike K will pay one unit of currency if $S_T \geq K$ at expiry, and nothing otherwise.

• A binary put will pay 1 if $S_T < K$, and nothing otherwise.

□

Let B_c be a binary call with strike K . The boundary value problem for B_c is

$$\begin{cases} \frac{\partial B_c}{\partial t} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 B_c}{\partial S^2} + rS \frac{\partial B_c}{\partial S} - rB_c = 0 \\ B_c(T) = I_{\{S_T \geq K\}} \end{cases}$$

Exercise 12.4.3 (1) The solution is of this BVP is given by

$$B_c(0) = e^{-rT} N(d_-)$$

where as before

$$d_- = \frac{\ln \frac{S_0}{K} + (r - \frac{1}{2}\sigma^2)T}{\sigma\sqrt{T}}$$

(2) You should be able to obtain a put–call parity for binary options, and then deduce the value of a binary put from that.

□

Chapter 13

Introduction to PDEs

In this chapter we examine what *partial differential equations* (PDEs) are, how they can be classified, and what is meant by a solution of a PDE.

13.1 What is a PDE?

An ODE is an equation involving a function of one variable and its derivatives. A PDE is an equation involving a function $u(x, y, \dots)$ of more than one variable, and its (partial) derivatives. They are used to model not only physical phenomena such as wave motion, heat conduction, fluid dynamics, electromagnetism etc., but are also used in the biological and economic sciences. In stochastic analysis, Markov processes are closely associated with PDEs.

Here are a few examples:

$u_x + u_y = 0$	transport eqn.	$u = u(t, x)$
$u_t = u_{xx}$	1D heat eqn.	$u = u(t, x)$
$u_t = u_{xx} + u_{yy} + u_{zz}$	3D heat eqn.	$u = u(t, x, y, z)$
$u_{xx} + u_{yy} = 0$	2D Laplace's eqn.	$u = u(x, y)$
$u_{tt} = u_{xx} + u_{yy} + u_{zz}$	3D wave eqn.	$u = u(t, x, y, z)$
$u_x + uu_y = 0$	shock wave	$u = u(x, y)$
$u_t + uu_x + u_{xxx} = 0$	dispersive wave	$u = u(t, x)$
$i\hbar u_t = -\frac{\hbar^2}{2m}u_{xx} + V(x)u$	1D Schrodinger eqn.	$u = u(t, x)$
$u_t + \frac{1}{2}\sigma^2 x^2 u_{xx} + rxu_x - ru = 0$	Black-Scholes eqn.	$u = u(t, x)$
$u_t = u_{xx} + ru(1 - u)$	Fisher's eqn.	$u = u(t, x)$

As stated before the *unknown function* u —the *dependent variable* — is always a function of *more than one independent variable* t, x, y, \dots . Loosely speaking, the dependent variable is the one that is being differentiated, whereas the independent variables are those we differentiate *with respect to*.

The general form of a PDE is thus:

$$F(u, u_{x_1}, \dots, u_{x_n}, u_{x_1 x_1}, u_{x_1 x_2}, \dots, u_{x_n x_n}, \dots, u_{x_i x_j x_k}, \dots, \text{other parameters}) = 0$$

13.1.1 Types of PDEs

PDEs can be classified according to the following criteria:

- **Order:** The order of a PDE is the order of the highest partial derivative in the equation. For example, the transport- and shock wave equations are first order, the heat-, wave-, Laplace- and Black-Scholes equations are second order, and the dispersive wave equation is third order.
- **Number of variables:** The number of independent variables.
- **Linearity:** We can write a PDE in *operator form* $\mathcal{L}u = f$, where f is a function of just the independent variables. For example, the 1D heat equation is

$$\mathcal{L}u = 0 \quad \text{where} \quad \mathcal{L} := \frac{\partial}{\partial t} - \frac{\partial^2}{\partial x^2}$$

The dispersive wave equation is

$$\mathcal{L}u = 0 \quad \text{where} \quad \mathcal{L} := \frac{\partial}{\partial t} + (\cdot) \frac{\partial^2}{\partial x^2} + \frac{\partial^3}{\partial x^3}$$

When the operator \mathcal{L} is *linear*, the PDE is said to be linear. This means that the *dependent variable and its derivatives occur linearly*. The independent variables may occur non-linearly. The heat equation is linear, whereas the dispersive wave equation is not — the latter has the non-linear term uu_{xx} . Here are some more examples:

$$u_x + \sin(y)u_{yy} = e^x \quad \text{linear} \quad u_{xx} + \sin(u) = 0 \quad \text{non-linear}$$

- **Homogeneity:** A PDE $\mathcal{L}u = f$ is *homogeneous* if $f = 0$ identically. Else it is *non-homogeneous*.
- **Kinds of coefficients:** The coefficients may be constant or variable.

13.2 Solutions to a PDE

13.2.1 Contrast with ODEs

ODEs usually have few types of solution. Consider, for example, the *first-order linear* ODE

$$y' + 2xy = 2x$$

which can be solved by using an *integrating factor*. The idea is to multiply both sides by a function $I(x)$ — the integrating factor — so that the lefthand side reduces to $(Iy)'$. Thus we want $Iy' + 2Ixy = (Iy)'$, from which it follows easily that $I' = 2Ix$. This yields a *separable* ODE for I , whose solution is

$$\frac{dI}{I} = 2x \, dx \implies \ln I = x^2 + c \implies I = Ce^{x^2}$$

We thus take $I(x) = e^{x^2}$ and obtain

$$(ye^{x^2})' = xe^{x^2}$$

so that

$$y = \frac{1}{2} + Ce^{-x^2} \quad C \text{ constant}$$

These are *all* the solutions of the ODE (i.e. one for each value of the constant C), and the ODE thus determines the nature of the solution.

The situation is dramatically different for PDEs. Consider the following first order linear PDE $u_x + u_y = 0$. Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be any differentiable function, and define $u(x, y) = f(x - y)$. Direct computation shows that

$$u_x + u_y = f'(x - y) - f'(x - y) = 0$$

so that $u = f(x - y)$ is a solution of the PDE. Hence each of the following is a solution of the given PDE:

$$\begin{aligned} u(x, y) &= 5x - 5y \\ u(x, y) &= (x - y)^2 \\ u(x, y) &= \frac{e^{x-y}}{(x - y)^3} \end{aligned}$$

etc. etc. etc. The nature of the solution is very far from being determined by the PDE, and we usually need more information. We shall shortly discuss what sort of additional information we require: *initial* or *boundary conditions*. But first, we make a short detour:

13.2.2 First-Order Linear PDEs

We would like to be able to solve at least a few PDEs before we continue our qualitative discussion of the solutions to PDEs, to build intuition and confidence. We therefore show, in this subsection, how to solve certain kinds of first-order linear PDEs. We begin by demonstrating a generalization of the technique used to solve the PDE $u_x + u_y = 0$ in the previous subsection:

Example 13.2.1 Consider the PDE

$$u_x + yu_y = 0$$

which is first-order linear and homogeneous. Note that

$$u_x + yu_y = (\partial_x u \quad \partial_y u) \begin{pmatrix} 1 \\ y \end{pmatrix} = \text{directional derivative of } u \text{ in direction } (1, y)$$

So the directional derivative of any solution in the direction $(1, y)$ is zero.

Consider now a function $y = y(x)$, which determines a curve consisting of points $(x, y(x))$ in the XY -plane. The tangent vector of such a curve at the point (x, y) is $(1, y'(x))$. Now the curves $y = y(x)$ which “point” in the direction $(1, y)$, i.e. which have tangent vector $(1, y)$ at the point (x, y) , and thus these curves satisfy $y' = y$. This ODE is easily solved: $y(x) = Ce^x$ where C is constant. The curves $y = Ce^x$ are called the *characteristic curves* of the PDE. On

each such curve, the directional derivative of a solution u is zero, and hence u is constant on the characteristic curves. Another way to see this:

$$\frac{du(x, Ce^x)}{dx} = u_x \cdot 1 + u_y \cdot Ce^x = u_x + yu_y = 0$$

It follows that

$$u(x, Ce^x) = u(0, Ce^0) = u(0, C) \quad \text{which is independent of } x$$

Given an arbitrary point (x, y) , we find C such that $y = Ce^x$, namely $C = e^{-x}y$. It then follows that

$$u(x, y) = u(0, e^{-x}y) \quad \text{which is a function of } e^{-x}y$$

Hence the general solution of the PDE is given by

$$u(x, y) = f(e^{-x}y) \quad \text{where } f : \mathbb{R} \rightarrow \mathbb{R} \text{ is differentiable}$$

□

Now try it yourself:

Exercise 13.2.2 Consider the PDE

$$u_x + 2xy^2u_y = 0$$

- (a) Show that the directional derivative of u at the point (x, y) in the direction $(1, 2xy^2)$ is zero.
- (b) Deduce that u is constant on the characteristic curves $y = \frac{1}{C-x^2}$, C constant.
- (c) Conclude that $u(x, y) = f(x^2 + \frac{1}{y})$

□

The above technique works for PDE's of the form $a(x, y)u_x + b(x, y)u_y = 0$. It reduces the solution of the PDE to that of an ODE $\frac{dy}{dx} = \frac{b(x, y)}{a(x, y)}$. The solutions of this ODE are the characteristic curves of the PDE, along which every solution is constant. We can extend it to homogeneous first-order linear PDEs with constant coefficients, as we now demonstrate by example:

Example 13.2.3 Consider the PDE

$$3u_x + 2u_y - 5u = 0$$

We would like to put it in the form $au_x + bu_y = 0$, i.e. we need to get rid of the u -term. Define a new function

$$v(x, y) = e^{-\frac{5}{3}x}u(x, y)$$

Then $v_x = e^{-\frac{5}{3}x}[u_x - \frac{5}{3}u]$ and so

$$3v_x + 2v_y = e^{-\frac{5}{3}x}[3u_x - 5u + 2u_y] = 0$$

We thus — this is an exercise! — solve the PDE $3v_x + 2v_y = 0$ as before to obtain $v(x, y) = f(y - \frac{2}{3}x)$, and hence the solution of the original PDE is

$$u(x, y) = e^{\frac{5}{3}x}f(y - \frac{2}{3}x)$$

as may easily be verified by direct differentiation.

□

Another technique for solving first-order linear involves *changing coordinates*:
Consider the PDE

$$au_x + bu_y + g(x, y)u = 0$$

Define new coordinates

$$\begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} = \begin{pmatrix} a & b \\ b & -a \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \quad \text{so that} \quad \begin{pmatrix} x \\ y \end{pmatrix} = \frac{1}{a^2 + b^2} \begin{pmatrix} a & b \\ b & -a \end{pmatrix} \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix}$$

Now by the chain rule for differentiation

$$u_x = au_{\bar{x}} + bu_{\bar{y}} \quad u_y = bu_{\bar{x}} - au_{\bar{y}}$$

so the PDE becomes

$$u_{\bar{x}} + \bar{g}(\bar{x}, \bar{y})u = 0 \quad \text{where} \quad \bar{g}(\bar{x}, \bar{y}) = \frac{1}{a^2 + b^2} g\left(\frac{a\bar{x} + b\bar{y}}{a^2 + b^2}, \frac{b\bar{x} - a\bar{y}}{a^2 + b^2}\right) = \frac{g(x, y)}{a^2 + b^2}$$

Thus the new PDE looks like a first-order linear ODE of the form $v' + f_y(x)v = 0$ with solution $v(x) = h(\bar{y})e^{-\int f_y(x) dx}$, where h is an arbitrary differentiable function of the “parameter” y only. Thus we get

$$u = h(\bar{y})e^{-\int \bar{g}(\bar{x}, \bar{y}) d\bar{x}}$$

which we can transform to (x, y) -coordinates to obtain u .

Example 13.2.4 We solve the PDE

$$u_x + 2u_y + (2x - y)u = 0$$

New coordinates are $\bar{x} = x + 2y$, $\bar{y} = 2x - y$, and the PDE becomes

$$u_{\bar{x}} + \frac{1}{5}\bar{y}u = 0$$

which can be solved to obtain

$$u = h(\bar{y})e^{-\frac{1}{5}\bar{x}\bar{y}} = h(2x - y)e^{-\frac{2x^2 + 3xy - 2y^2}{5}} \quad h \text{ arbitrary}$$

□

Example 13.2.5 We solve the PDE

$$u_x + 2u_y + (2x - y)u = 2x^2 + 3xy - 2y^2$$

We have already solved the homogeneous PDE $u_x + 2u_y + (2x - y)u = 0$ in the previous example, with general solution $u(x, y) = h(2x - y)e^{-\frac{2x^2 + 3xy - 2y^2}{5}}$. When we move to the (\bar{x}, \bar{y}) -coordinates described there, the PDE becomes

$$u_{\bar{x}} + \frac{1}{5}\bar{y}u = \frac{1}{5}\bar{x}\bar{y}$$

This looks like a first-order linear ODE involving a “parameter” \bar{y} , with integrating factor $I(\bar{x}) = e^{\frac{\bar{y}\bar{x}}{5}}$, and yields

$$\frac{dIu}{d\bar{x}} = \bar{x}e^{C\bar{x}} \quad \text{where} \quad C = \frac{\bar{y}}{5}$$

Integrating by parts, we obtain

$$Iu = \bar{x}e^{C\bar{x}} - \frac{1}{C}e^{C\bar{x}} + D$$

where the “constant” of integration D may involve the “parameter” \bar{y} . Thus

$$u = \bar{x} - \frac{1}{C} + De^{-C\bar{x}}$$

Translating back to (x, y) -coordinates, we see that

$$u(x, y) = x + 2y - \frac{5}{2x-y} + D(2x-y)e^{-\frac{2x^2+3xy-2y^2}{5}} \quad D \text{ arbitrary}$$

□

Changes of coordinates will play an important role in the next section, where we take a first look at second-order linear PDEs.

We have now seen that the characteristic curves of a first-order linear PDE are related to a choice of coordinates which simplifies the PDE, and turns it into an ODE. The following exercise gives another important interpretation: The characteristic curves of a PDE $L[u] = f$ are exactly those curves along which knowledge of the values of u does not determine the values of u_x, u_y . We will return to this when we discuss characteristics for second-order linear PDEs.

Exercise 13.2.6 We show that the characteristic curves of a *quasilinear* first-order PDE

$$a(x, y, u)u_x + b(x, y, u)u_y = c(x, y, u) \quad (*)$$

are exactly those curves along which the PDE $(*)$ together with knowledge of the values u along the curve is insufficient to determine u_x, u_y .

6.1 Let Γ be a curve in the xy -plane, parametrized by a variable t :

$$x = x(t) \quad y = y(t) \quad \text{along } \Gamma$$

u is supposedly known along Γ , so there is a known function f such that

$$u(x(t), y(t)) = f(t)$$

Show that we have the following system of linear equations for u_x, u_y :

$$\begin{aligned} x'u_x + y'u_y &= f' \\ au_x + bu_y &= c \end{aligned}$$

where $a = a(x, y, u) = a(x(t), y(t), f(t))$ is known, and the same is true for b, c .

6.2 Show that this system uniquely determine u_x, u_y , unless

$$x'b - y'a = 0 \quad (\dagger)$$

6.3 Show that (\dagger) holds precisely when $b \, dx - a \, dy = 0$ along Γ , i.e. precisely when Γ is a characteristic of $(*)$.

□

We end our detour with one more example, after which we will continue our qualitative discussion of solutions of PDEs.

Example 13.2.7 The one-dimensional *wave* operator (or D'Alembert operator) $\frac{\partial^2}{\partial t^2} - c^2 \frac{\partial^2}{\partial x^2}$ factorizes:

$$\partial_{tt}^2 - c^2 \partial_{xx}^2 = (\partial_t + c\partial_x)(\partial_t - c\partial_x)$$

As a result, it is easy to find the general solution of the wave equation $u_{tt} - c^2 u_{xx} = 0$: If u is a solution, then

$$v := (\partial_t - c\partial_x)u \quad \text{is a solution to} \quad v_t + cv_x = 0$$

This is a homogeneous first order linear equation, with general solution $v = h(x - ct)$, where h is an arbitrary function. We thus obtain $(\partial_t - c\partial_x)u = v$, which is an *inhomogeneous* first order linear equation. The corresponding *homogeneous* equation $u_t - cu_x = 0$ has general solution $g(x + ct)$, where g is arbitrary. If we try to find a *particular solution* of the form $u_1(x, t) = f(x - ct)$ to the inhomogeneous equation $u_t - cu_x = v$, we find that

$$f'(s) = \frac{h(s)}{2c} \quad \text{and so} \quad f(s) = \frac{1}{2c} \int h(s) \, ds$$

But since h, f is arbitrary, i.e.

$$u_1(x, t) = f(x + ct) \quad f \text{ an arbitrary function}$$

If u is another solution to the inhomogeneous equation $u_t - cu_x = h$, then (by linearity) $u - u_1$ is a solution to the homogeneous equation, i.e. $u - u_1 = g(x + ct)$, where g is arbitrary. It follows that

$$u(t, x) = f(x + ct) + g(x - ct)$$

where f, g are arbitrary. This should be verified by direct differentiation.

Note that $f(x + ct)$ can be interpreted as the graph of $f(x)$ moving to the left at speed c , whereas $g(x - ct)$ is the graph of $g(x)$ moving to the right at speed c . The parameter c in the wave equation $u_{tt} - c^2 u_{xx}$ can therefore be interpreted as the speed of the wave.

□

13.2.3 Initial- and Boundary Conditions

There is no general mathematical theory for solving partial differential equations, and one can seldom solve the PDE in general, as we were able to do for the first-order PDEs and one-dimensional wave equation in the preceding subsection. Instead, research has focussed on particular classes of PDEs, often suggested by an applied/physical problem. Different techniques have developed for solving different classes of PDEs. And certain auxiliary conditions, also suggested by the applied/physical problem, are imposed as well. These are of the following types:

Initial Value Problems (IVP) As suggested by the name, these specify the state at an initial time, usually $t = 0$. IVP are also called *Cauchy problems*.

Examples 13.2.8 1. Consider the one-dimensional heat equation

$$u_t = ku_{xx} \quad k > 0$$

which governs the evolution of the temperature in an infinitely long homogeneous rod: $u(t, x)$ is the temperature at point x at time t . This PDE has many solutions. But using physical intuition about temperatures, one would guess that the temperature at times $t > 0$ is completely determined by the temperature at $t = 0$, i.e. that there is *just one* function u which satisfies the IVP

$$\left. \begin{array}{ll} u_t = ku_{xx} & \text{PDE} \\ u(0, x) = \Phi(x) & \text{Initial Condition} \end{array} \right\} = \text{Initial Value Problem}$$

2. Consider the one-dimensional wave equation

$$u_{tt} = c^2 u_{xx}$$

which governs the evolution of a vibrating string: $u(t, x)$ is the displacement at point x at time t . Physical intuition suggests that we need two initial conditions to specify the solution uniquely — not just the initial displacement, but also the initial velocity:

$$\left. \begin{array}{lll} u_{tt} = c^2 u_{xx} & \text{PDE} & + \\ u(0, x) = \Phi(x) & \text{IC} & + \\ u_t(0, x) = \Psi(x) & \text{IC} & \end{array} \right\} = \text{IVP}$$

In these examples, we see from physical grounds that we need different kinds of data to find unique solutions to these PDEs. (This must, of course, also be proven mathematically.)

□

Boundary Value Problems (BVP) Given a particular applied problem modelled by a PDE, the PDE is usually only valid in a certain open set U , and the nature of the problem may require that the solution u satisfy certain auxiliary conditions on the boundary ∂U of U . In practice, these boundary conditions are of three types:

I. *Dirichlet conditions* specify the value of the solution u on the boundary:

$$u = f \quad \text{on } \partial U$$

II. *Neumann conditions* specify the value of the outward normal derivative $\frac{\partial u}{\partial n} := Du \cdot \mathbf{n}$ on the boundary ∂U . (Here, \mathbf{n} is a generic unit outward pointing vector which is normal to the boundary ∂U .)

$$\frac{\partial u}{\partial n} = f \quad \text{on } \partial U$$

III. *Robin conditions* specify the value of $\frac{\partial u}{\partial n} + au$ on the boundary ∂U . Here, a is a function of the independent variables t, x, y, z, \dots

$$\frac{\partial u}{\partial n} + au = f \quad \text{on } \partial U$$

Examples 13.2.9 1. Consider a vibrating string of length L . The appropriate PDE, the one-dimensional wave equation $u_{tt} = c^2 u_{xx}$ is valid on the open interval $U = (0, L)$, whose boundary consists of just the two endpoints: $\partial U = \{0, L\}$.

- If both ends of the string are held fixed, then we have homogeneous Dirichlet boundary conditions:

$$u(t, 0) = 0 \quad u(t, L) = 0 \quad \text{all } t \geq 0$$

- Suppose now that the $x = 0$ end is fixed, but that the $x = L$ end is able to freely slide up and down along a frictionless track. Then we have

$$u(t, 0) = 0 \quad u_x(t, L) = 0 \quad \text{all } t \geq 0$$

i.e. a Dirichlet condition holds on the left end, and a Neumann condition on the right — Note that the outward normal derivative $\frac{\partial u}{\partial n}$ is just the ordinary partial derivative $\frac{\partial u}{\partial x}$.

- If the $x = 0$ end is fixed and the $x = L$ end is able to move up and down along a frictionless track, but is attached to a spring or rubber band (satisfying Hooke's Law), then

$$u(t, 0) = 0 \quad u_x(t, L) + ku = 0 \quad \text{all } t \geq 0$$

i.e. a Dirichlet condition on the left boundary, and a Robin condition on the right.

2. Suppose that an object occupies a region of space whose interior is the open set U . The evolution of the temperature is governed by the heat equation $u_t = u_{xx} + u_{yy} + u_{zz}$.

- If the object is immersed in a heat reservoir of temperature $g(t)$, we would have Dirichlet boundary conditions: $u(t, \mathbf{x}) = g(t)$ for $\mathbf{x} \in \partial U$.
- If the object is perfectly insulated, then no heat flows across the boundaries, i.e. we have Robin boundary conditions: $\frac{\partial u}{\partial n} = 0$.
- If the object is immersed in a medium at temperature T_0 , then an inhomogeneous Robin boundary condition would hold at that end: $\frac{\partial u}{\partial n} = -k(u - T_0)$.

□

Initial–Boundary Value Problems (IBVP) Many problems involve both initial and boundary conditions. For example, if a rod of length L is insulated along its length, has an initial temperature $\Phi(x)$, has the $x = 0$ -end kept at a temperature $g(t)$ and the $x = L$ -end immersed in a reservoir of temperature T_0 , then we have the IBVP

$$\left. \begin{aligned} u_t &= u_{xx} \\ u(0, x) &= \Phi(x) \\ u(t, 0) &= g(t) \\ u_x(t, L) &= -k(u(L, t) - T_0) \end{aligned} \right\}$$

13.2.4 Well-Posed Problems

As said before, PDEs can have many solutions. Consider the following example:

Example 13.2.10 Consider a homogeneous rod of length L and let $u(t, x)$ be the temperature at time t at point x . The $x = 0$ end of the rod is kept at 0°C and the $x = L$ end at 100°C . Suppose now that the rod has reached equilibrium, i.e. that the temperature is not changing over time. Then $u(t, x) =: U(x)$ is a function of just x . To determine the temperature $U(x)$ of the rod at point x in equilibrium we must solve the following BVP: The heat equation $u_t = ku_{xx}$ becomes $0 = U_{xx}$, and the boundary conditions become $U(0) = 0, U(L) = 100$. Thus

$$U_{xx} = 0 \text{ on } (0, L) \quad \text{subject to} \quad U(0) = 0 \quad U(L) = 100$$

Now

$$U_1(x) := \begin{cases} 5 & \text{if } 0 < x < L \\ 0 & \text{if } x = 0 \\ 100 & \text{if } x = L \end{cases}$$

is clearly a solution to the BVP. So is

$$U_2(x) := \begin{cases} 20x + 72 & \text{if } 0 < x < L \\ 0 & \text{if } x = 0 \\ 100 & \text{if } x = L \end{cases}$$

Yet physical intuition tells us that *the* correct solution is

$$U(x) = \frac{100x}{L}$$

U_1, U_2 are pathological: We have simply taken arbitrary solutions of the PDE and pasted the boundary conditions onto them. Note that U_1, U_2 are continuous on $(0, L)$ but not on $[0, L]$. U , however, *is* continuous on $[0, L]$. Moreover, it is easy to see that this is the *only* solution which is continuous on $[0, L]$. Thus by demanding, for this problem, that the solution is continuous to the boundaries, we guarantee a *unique solution*.

This suggests that we should require some regularity in our solutions.

□

An initial and/or boundary value problem is said to be **well-posed** if and only if:

- (i) *Existence*: There is at least one solution.
- (ii) *Uniqueness*: There should be no more than one solution.
- (iii) *Continuous dependence on the data*: Small changes in the data produce only small changes in the solution.

(i) and (ii) allow us to talk about *the* solution (as opposed to *a* solution), and (iii) is a requirement if we want to be able to calculate the solution: Calculation often involves approximation, e.g. in terms of limiting processes such as integrals, series expansions, etc.

Related to (i)-(iii) above are three “practical” questions:

- (i') How do we *construct* a solution?
- (ii') How *regular* (e.g. continuous, differentiable) is the solution?
- (iii') If an exact analytic formula for a solution cannot be found, how do we approximate the solution?

Remarks 13.2.11 The context of the problem determines what kind of extra regularity conditions we should impose on the problem. In the example of the rod in thermal equilibrium, above, we imposed continuity on the boundaries to obtain a unique solution. One might think that, e.g. we should always require that an n^{th} order PDE be n -times differentiable on the interior of its domain. This is not true, as PDEs can also be used to model inherently discontinuous phenomena, such as shock waves, and mathematicians have come up with various kinds of objects that may count as a solution – not only n -times differentiable functions — by interpreting the PDE for these objects in different ways. The bottom line is that the context of the PDE (i.e. the problem the PDE is trying to model) is always important.

□

For a PDE together with some auxiliary conditions, such as initial- and/or boundary conditions, to be well-posed, there must be a delicate balance. There can't be too few auxiliary conditions, or the solution will not be unique (e.g. the vibrating string where the initial displacement is given, but not the initial velocity). There can't be too many auxiliary conditions, or they will “clash” (i.e. be mutually inconsistent), and there won't be a solution. We will discuss the well-posedness of certain classes of PDEs later in the course.

13.3 Classification of Linear Second-Order PDEs

It is an interesting fact that many phenomena may be modelled — at least as a first approximation — by *linear second order PDEs*. The heat, wave and Laplace equations are canonical examples of 2nd order linear PDEs. The Laplace operator $\Delta = \frac{\partial^2}{\partial x_1^2} + \cdots + \frac{\partial^2}{\partial x_n^2}$ plays an important role. We recall the two-dimensional versions, with $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$:

$$\begin{aligned} \text{Heat equation :} & \quad \frac{\partial u}{\partial t} - \Delta u = 0 \\ \text{Wave equation :} & \quad \frac{\partial^2 u}{\partial t^2} - \Delta u = 0 \\ \text{Laplace's equation :} & \quad \Delta u = 0 \end{aligned}$$

The general second order linear PDE has the form

$$\underbrace{\sum_{i,j=1}^n a_{ij} u_{x_i x_j}}_{\text{principal part}} + \sum_{i=1}^n b_i u_{x_i} + cu = d$$

where the a_{ij}, b_i, c, d are functions of the independent variables x_i only. Because $u_{x_i x_j} = u_{x_j x_i}$, the coefficients a_{ij} of the principal part may always be arranged so that $a_{ij} = a_{ji}$, i.e. so that the matrix $A := (a_{ij})$ is symmetric. The linear second order differential equations may be further classified into three groups, according to the properties of the matrix A .

Recall the following spectral decomposition theorem from linear algebra:

Theorem 13.3.1 *Every symmetric matrix has an orthonormal basis of eigenvectors.*

□

Remarks 13.3.2 1. Suppose that A is symmetric, with orthonormal basis of eigenvectors \mathbf{v}_i , so that $A\mathbf{v}_i = \lambda_i\mathbf{v}_i$ for eigenvalues λ_i . Consider the matrix O whose columns are the eigenvectors \mathbf{v}_i , i.e. $O_{ij} = v_{ji}$. Then

$$(O^{tr}O)_{ij} = \sum_k O_{ik}^{tr}O_{kj} = \sum_k v_{ik}v_{jk} = \langle \mathbf{v}_i, \mathbf{v}_j \rangle = \delta_{ij}$$

i.e. $O^{tr} = O^{-1}$. Now $AO = OD[\lambda]$, where $D[\lambda]$ is the diagonal matrix with eigenvalues along the diagonal, i.e. $D[\lambda]_{ij} = \delta_{ij}\lambda_i$. It follows that

$$O^{tr}AO = D[\lambda]$$

i.e. that A is diagonalizable.

2. Now recall that a symmetric matrix is said to be non-negative definite (or positive semidefinite) if and only if

$$\mathbf{x}^{tr}A\mathbf{x} \geq 0 \quad \text{for all } \mathbf{x}$$

This will be the case precisely when all A 's eigenvalues are non-negative. For if $\lambda_i \geq 0$ for all i , and if $\mathbf{x} = \sum_i \alpha_i \mathbf{v}_i$ is arbitrary, then

$$\mathbf{x}^{tr}A\mathbf{x} = \left(\sum_i \alpha_i \mathbf{v}_i\right)^{tr} A \left(\sum_j \alpha_j \mathbf{v}_j\right) = \sum_{i,j} \alpha_i \alpha_j \lambda_j \delta_{ij} = \sum_i \alpha_i^2 \lambda_i \geq 0$$

Conversely, if A is non-negative definite, then

$$0 \leq \mathbf{v}_i^{tr} A \mathbf{v}_i = \lambda_i \|\mathbf{v}_i\|^2$$

so that each eigenvalue λ_i is ≥ 0 .

Similar remarks can be made for strictly positive definite and for negative definite matrices.

3. An orthogonal transformation $B = O^{tr}AO$ does not change the “definiteness” of a matrix, because B and A have the same eigenvalues: If \mathbf{v} is an eigenvector of A with corresponding eigenvalue λ , then $O^{tr}\mathbf{v}$ is an eigenvector of B , with the same eigenvalue. (Note that B is automatically symmetric if A is.)

□

Suppose we have a 2nd order linear PDE

$$\sum_{i,j=1}^n a_{ij}u_{x_i x_j} + \sum_{i=1}^n b_i u_{x_i} + cu = d$$

Define new variables ξ_r ($r = 1, \dots, n$) by a linear change as follows:

$$\xi_r = \sum_i p_{ir} x_i \quad \text{i.e.} \quad \xi = P^{tr} \mathbf{x} \quad \text{where} \quad P = (p_{ir})$$

Then

$$\frac{\partial}{\partial x_i} = \sum_r \frac{\partial \xi_r}{\partial x_i} \frac{\partial}{\partial \xi_r} = \sum_r p_{ir} \frac{\partial}{\partial \xi_r}$$

and similarly

$$\frac{\partial^2}{\partial x_i \partial x_j} = \sum_{r,s} p_{ir} p_{js} \frac{\partial^2}{\partial \xi_r \partial \xi_s}$$

In the new coordinates, the principal part of the given PDE therefore takes the form

$$\begin{aligned} \sum_{i,j} a_{ij} u_{x_i x_j} &= \sum_{i,j} a_{ij} \sum_{r,s} p_{ir} p_{js} u_{\xi_r \xi_s} \\ &= \sum_{r,s} \left[\sum_{i,j} p_{ir} a_{ij} p_{js} \right] u_{\xi_r \xi_s} \\ &= \sum_{r,s} c_{rs} u_{\xi_r \xi_s} \end{aligned}$$

i.e. the principal part has matrix $C = P^{tr} A P$.

By ensuring that C is a diagonal matrix, we can remove the mixed partial derivatives $u_{\xi_r \xi_s}$, $r \neq s$. In particular, it follows from the preceding remarks that if we take $P = O$, the matrix whose columns form an orthonormal basis of eigenvectors, then $C = O^{tr} A O = D[\lambda]$, i.e.

$$\sum_{r,s} c_{rs} u_{\xi_r \xi_s} = \sum_r \lambda_r u_{\xi_r \xi_r}$$

Hence any second order linear differential equation can have its principal part reduced to diagonal form — i.e. involving no mixed partial derivatives — by a linear change of variables.

We now classify the 2nd order linear PDEs as follows:

- I. If A is strictly positive definite or strictly negative definite, i.e. if all the eigenvalues are non-zero and have the same sign, we say that the PDE is **elliptic**.

Laplace's equation is the canonical example of an elliptic PDE.

- II. If A has a zero eigenvalue, i.e. if $\det(A) = 0$, then the PDE is **parabolic**.

The heat equation is the canonical example of a parabolic equation.

- III. If all the eigenvalues are non-zero and one has sign opposite to all the others, we say the PDE is **hyperbolic**.

The wave equation is the canonical example of a hyperbolic equation.

- IV. Any other type is called ultrahyperbolic — but these do not occur often in practice, and only in dimension $n \geq 4$.

□

In stochastic analysis and finance, it is the second order parabolic equations which occur most frequently.

Exercise 13.3.3 Consider a second order linear PDE

$$Au_{xx} + 2Bu_{xy} + Cu_{yy} + Du_x + Eu_y + F = G$$

Show that the PDE is

- (a) Elliptic if $B^2 - AC < 0$;
- (b) Parabolic if $B^2 - AC = 0$;
- (c) Hyperbolic if $B^2 - AC > 0$.

Remarks 13.3.4 The names *elliptic*, *parabolic* and *hyperbolic* are adopted from conic sections. For example, an ellipse has basic equation $ax^2 + by^2 = \text{const.}$, with $a, b > 0$, and this can always be reduced to $\bar{x}^2 + \bar{y}^2 = \text{const.}$ by a change of variables — just take $\bar{x} := \sqrt{a}x$, $\bar{y} := \sqrt{b}y$. So after a change of variables, an ellipse becomes a circle. Now the expression $x^2 + y^2$ is formally analogous to $u_{xx} + u_{yy}$. Indeed

$$x^2 + y^2 = \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \quad \text{whereas} \quad u_{xx} + u_{yy} = \begin{pmatrix} \partial_x & \partial_y \end{pmatrix} \begin{pmatrix} \partial_x \\ \partial_y \end{pmatrix} u$$

So Laplace's equation $u_{xx} + u_{yy} = 0$ is said to be elliptic.

Similarly, the basic equation of a parabola is $y - x^2 = \text{const.}$, and hence the heat equation $u_y - u_{xx} = 0$ is said to be parabolic. Hyperbolas have two forms for the basic equation: $xy = \text{const.}$ and $x^2 - y^2 = \text{const.}$ (If you are only familiar with the first form, note that the change of variables $\bar{x} = \frac{x+y}{2}$, $\bar{y} = \frac{x-y}{2}$ transforms $xy = \text{const.}$ to $\bar{x}^2 - \bar{y}^2 = \text{const.}$) Hence the PDEs $u_{xy} = 0$ and $u_{xx} - u_{yy} = 0$, both forms of the wave equation, are said to be hyperbolic.

□

This analogy between second order linear PDEs with constant coefficients and conic sections carries over to more complicated situations:

Example 13.3.5 Consider the the PDE

$$9u_{xx} - 24u_{xy} + 16u_{yy} - 18u_x - 101u_y + 19 = 0$$

An inspection of the principal part shows that it is parabolic: $b^2 - ac = 12^2 - 9 \times 16 = 0$. We show how it can be transformed to a heat equation by a change of variables, and simultaneously show the strength of the connection with conic sections. The principal part is

$$9u_{xx} - 24u_{xy} + 16u_{yy} = \begin{pmatrix} \partial_x & \partial_y \end{pmatrix} \begin{pmatrix} 9 & -12 \\ -12 & 16 \end{pmatrix} \begin{pmatrix} \partial_x \\ \partial_y \end{pmatrix} u$$

The principal part of the corresponding conic section is

$$9x^2 - 24xy + 16y^2 = \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} 9 & -12 \\ -12 & 16 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

The eigenvalues of the matrix A which determines this *quadratic form* are easily determined to be $\lambda = 0$ and $\lambda = 25$, with corresponding eigenvectors $\begin{pmatrix} 4 \\ 3 \end{pmatrix}^{tr}$ and $\begin{pmatrix} -3 \\ 4 \end{pmatrix}^{tr}$, both normalized

to have unit length. The orthogonal matrix O whose columns are these eigenvectors is used to define new coordinates:

$$\begin{pmatrix} \xi \\ \eta \end{pmatrix} = O^{tr} \begin{pmatrix} x \\ y \end{pmatrix} \quad \text{so that} \quad \begin{pmatrix} x \\ y \end{pmatrix} = O \begin{pmatrix} \xi \\ \eta \end{pmatrix} \quad \text{where} \quad O = \begin{pmatrix} \frac{4}{5} & -\frac{3}{5} \\ \frac{3}{5} & \frac{4}{5} \end{pmatrix}$$

The equation of the conic section in the new coordinates is therefore

$$\begin{pmatrix} \xi & \eta \end{pmatrix} O^{tr} A O \begin{pmatrix} \xi \\ \eta \end{pmatrix} - 18(4/5 \quad -3/5) \begin{pmatrix} \xi \\ \eta \end{pmatrix} - 101(3/5 \quad 4/5) \begin{pmatrix} \xi \\ \eta \end{pmatrix} + 19 = 0$$

i.e.

$$25\eta^2 - 75\xi - 70\eta + 19 = 0$$

Completing the square shows that this is equivalent to

$$\left(\eta - \frac{7}{5}\right)^2 = 3\left(\xi + \frac{2}{5}\right)$$

and thus we see that in (ξ, η) -coordinates the conic section is a parabola. However, the (ξ, η) -coordinates were obtained from the (x, y) -coordinates via a orthogonal matrix, i.e. a rotation. Since a rotated parabola is still a parabola, we see that the conic section $9x^2 - 24x + 16y^2 - 18x - 101y + 19 = 0$ defines a parabola.

We can further simplify the equation of the parabola to the form by a translation of the axes: We get

$$r^2 - t = 0 \quad \text{where} \quad t := 3\left(\xi + \frac{2}{5}\right), \quad r := \eta - \frac{7}{5}$$

Repeating all these steps for the PDE, as described before when we discussed removal of mixed partial derivatives, we see that in (ξ, η) -coordinates the PDE becomes

$$25u_{\eta\eta} - 75u_{\xi} - 70u_{\eta} + 19 = 0$$

which is equivalent to the heat equation when we move to (r, t) -coordinates

$$u_{rr} - u_t = 0$$

□

Exercises 13.3.6 Classify and transform the following PDEs so that the principal part is in canonical form:

1. $u_{xx} - u_{xt} - u_{tt} - u_x - u_t = 0$
2. $2u_{xx} + u_{xt} + 2u_{tt} + u = 0$
3. $-3u_{xx} + 10u_{xt} - 3u_{tt} + 7u_x - u_t + u = 0$

□

The above technique shows how to reduce the principal part of a second-order linear PDE with constant coefficients to canonical form. A simple trick allows us to remove the lower order derivatives:

Exercise 13.3.7 (a) Given

$$u_{x_1x_1} + u_{x_2x_2} - u_{x_3x_3} + 6u_{x_1} - 14u_{x_2} + 8u_{x_3} = 0$$

show that a transformation of dependent variable

$$u(x_1, x_2, x_3) = v(x_1, x_2, x_3)e^{-\sum_i c_i x_i}$$

transforms the PDE to

$$v_{x_1x_1} + v_{x_2x_2} - v_{x_3x_3} + (6+2c_1)v_{x_1} + (-14+2c_2)v_{x_2} + (8+2c_3)v_{x_3} + (c_1^2 + c_2^2 - c_3^2 + 6c_1 - 14c_2 + 8c_3)v = 0$$

Now choose $c_1 = -3, c_2 = 7, c_3 = -4$ to obtain

$$v_{x_1x_1} + v_{x_2x_2} - v_{x_3x_3} + 126v = 0$$

(b) Find a change of dependent variable that reduces the parabolic PDE

$$u_{xx} + 4u_x - 2u_t + 8u = 0$$

to the one-dimensional heat equation $v_t = \kappa v_{xx}$.

[Hint: Define v by $u(t, x) = v(t, x)e^{ax+bt}$.]

□

From the above examples and exercises it is clear that any second order linear PDE with constant coefficients can, via a transformation of coordinates and/or change of dependent variable, be reduced to either the Laplace-, or the heat- or the wave equation, depending on whether it is elliptic, parabolic or hyperbolic. We have perhaps spent an inordinate amount of time on this, and will have still more to say about it in the next section. It is therefore important that you understand why we emphasize this classification of PDEs: Essentially, there are just three kinds of behaviour a solution to a second order linear PDE can exhibit. The solution can either imitate a system in equilibrium (if it is elliptic), or, for evolving systems, it can imitate something like heat diffusion (if it is parabolic) or something like wave motion (if it is hyperbolic). After all, a smooth change of coordinates shouldn't change the qualitative properties of the solution, and a solution which behaves like diffusion in one set of coordinates will presumably behave like diffusion in another set of coordinates as well. The classification of 2nd order linear PDEs thus allows you to use your intuition about physical processes to obtain qualitative information about the nature of the solutions of these PDEs.

13.4 Characteristics for 2nd-Order Linear PDEs

When one seeks to reduce a second order linear PDE to canonical form, the notion of *characteristic curves* (or just *characteristics*) is often useful. The characteristics also have an interpretation as curves along which information can propagate, as we will endeavour to show later.

Consider a 2nd order linear PDE with principal part $au_{xx} + 2bu_{xy} + cu_{yy}$ in new coordinates $\xi = \xi(x, y), \eta = \eta(x, y)$. We assume that the change is invertible, or equivalently, by the Implicit Function Theorem, that $\xi_x\eta_y - \xi_y\eta_x \neq 0$. Do the following exercise now:

Exercise 13.4.1 (a) Recall the chain rule and show that

$$\begin{aligned} u_{xx} &= u_{\xi\xi}\xi_x^2 + 2u_{\xi\eta}\xi_x\eta_x + u_{\eta\eta}\eta_x^2 + u_{\xi}\xi_{xx} + u_{\eta}\eta_{xx} \\ u_{yy} &= u_{\xi\xi}\xi_y^2 + 2u_{\xi\eta}\xi_y\eta_y + u_{\eta\eta}\eta_y^2 + u_{\xi}\xi_{yy} + u_{\eta}\eta_{yy} \\ u_{xy} &= u_{\xi\xi}\xi_x\xi_y + u_{\xi\eta}(\xi_x\eta_y + \xi_y\eta_x) + u_{\eta\eta}\eta_x\eta_y + u_{\xi}\xi_{xy} + u_{\eta}\eta_{xy} \end{aligned}$$

(b) Hence show that the principal part of the PDE in transformed coordinates is

$$\tilde{a}u_{\xi\xi} + 2\tilde{b}u_{\xi\eta} + \tilde{c}u_{\eta\eta}$$

where

$$\tilde{a} = a\xi_x^2 + 2b\xi_x\xi_y + c\xi_y^2 \quad \tilde{c} = a\eta_x^2 + 2b\eta_x\eta_y + c\eta_y^2$$

and

$$\tilde{b} = a\xi_x\eta_x + b(\xi_x\eta_y + \xi_y\eta_x) + c\xi_y\eta_y$$

(c) Then show that

$$\tilde{b}^2 - \tilde{a}\tilde{c} = (b^2 - ac)(\xi_x\eta_y - \xi_y\eta_x)^2$$

so that

$$\tilde{A} = J^{tr} A J \quad \text{where} \quad J = \begin{pmatrix} \xi_x & \eta_x \\ \xi_y & \eta_y \end{pmatrix}$$

i.e. J is the Jacobian matrix of the map $(\xi, \eta) : \mathbb{R}^2 \rightarrow \mathbb{R}^2$. Thus that the type (elliptic, parabolic or hyperbolic) of a 2nd order linear PDE doesn't change under an invertible change of variables, linear or non-linear.

□

The principal part will simplify to just a $u_{\xi\eta}$ -term when $\tilde{a} = 0 = \tilde{c}$, i.e. when both ξ, η solve the equation

$$\begin{pmatrix} z_x & z_y \end{pmatrix} A \begin{pmatrix} z_x \\ z_y \end{pmatrix} = 0 \quad (\star)$$

This is the *characteristic equation* of the PDE. The level curves $z(x, y) = \text{const.}$ are known as the *characteristics* of the PDE. To determine its solutions, we use the following result, which reduces to an ODE:

Theorem 13.4.2 $z(x, y) = \text{const.}$ is a characteristic iff $z(x, y) = \text{const.}$ is a solution of the equation

$$a(dy)^2 - 2b(dx) \cdot (dy) + c(dx)^2 = 0 \quad (\dagger)$$

Proof: Suppose first that $z(x, y) = \gamma = \text{const.}$ is a characteristic curve, i.e. a solution of (\dagger) . If (\dagger) is not vacuous, then either $z_x(x, y)$ or $z_y(x, y)$ is $\neq 0$. Assume $z_y(x, y) \neq 0$, so that the equation $z(x, y) = \gamma$ defines y in terms of x , i.e.

$$z(x, y) = \gamma \quad \Leftrightarrow \quad y = f_\gamma(x)$$

Then along the curve $y = f_\gamma(x)$ we have $dz = 0$, and so

$$\frac{dy}{dx} = -\frac{z_x(x, y)}{z_y(x, y)}$$

From (†), upon dividing each term by z_y^2 , it now follows that

$$a\left(\frac{dy}{dx}\right)^2 - 2b\frac{dy}{dx} + c = 0$$

with (★) as immediate consequence.

A similar argument holds if $z_y = 0$ but $z_x \neq 0$.

Conversely, suppose that $z(x, y) = \text{const.}$ is a general solution of (†). We want to show that $z(x, y)$ satisfies (★) at an arbitrary point. Given an arbitrary point (x_0, y_0) , define $\gamma_0 = z(x_0, y_0)$, and consider the curve $y = f_{\gamma_0}(x)$. Along this curve, we have

$$0 = a\left(\frac{dy}{dx}\right)^2 - 2b\frac{dy}{dx} + c = a\left(\frac{z_x}{z_y}\right)^2 - 2b\left(-\frac{z_x}{z_y}\right) + c$$

In particular, with $x = x_0$ (and thus $y = f_{\gamma_0}(x_0) = y_0$) we see that (★) holds at x_0, y_0 .

—

Examples 13.4.3 1. The characteristics of the 1D wave equation $u_{tt} - \alpha^2 u_{xx} = 0$ are determined as follows: The characteristic equation is $z_t^2 - \alpha^2 z_x^2 = 0$, i.e. $a = 1, b = 0, c = -\alpha^2$ and so the characteristic curves are the solutions of $(dx)^2 - \alpha^2(dt)^2 = 0$. Hence $\frac{dx}{dt} = \pm 1$, i.e. $x = \pm \alpha t + \text{const.}$. So there are two families of characteristic curves, given by $\xi = x + \alpha t = \text{const.}$ and $\eta = x - \alpha t = \text{const.}$

2. The characteristics of the 1D heat equation $u_t - u_{xx} = 0$ are determined as follows: We have $a = 0 = b$ and $c = -1$, so the characteristic curves are solutions of $(dt)^2 = 0$, i.e. $t = \text{const.}$. There is just one family of characteristic curves, given by $\eta = t = \text{const.}$

3. The characteristic curves of the 2D Laplace equation $u_{xx} + u_{yy} = 0$ are determined as follows: We have $a = 1 = c$ and $b = 0$, so the characteristic curves are solutions of $(dy)^2 + (dx)^2 = 0$ which has no non-constant (real) solution. Hence Laplace's equation has no (real) characteristics.

□

Examples 13.4.4 1. **Problem:** Transform the equation $u_{xx} - x^2 y u_{yy} = 0$, ($y > 0$) to a canonical form with principal part $u_{\xi\eta}$.

Solution: We have $b^2 - ac = x^2 y > 0$ when $y > 0$ so the equation is hyperbolic. We first determine the characteristics, i.e. the solutions of $(dy)^2 - x^2 y (dx)^2 = 0$. This yields an ODE

$$\frac{dy}{dx} = \pm x\sqrt{y}$$

which is separable: $\frac{dy}{\sqrt{y}} = \pm x dx$, and hence $2\sqrt{y} = \frac{1}{2}x^2 + \text{const.}$, which yields

$$x^2 \pm 4\sqrt{y} = \text{const.}$$

We therefore define

$$\xi = x^2 + 4\sqrt{y} \quad \eta = x^2 - 4\sqrt{y}$$

Then, since $\tilde{A} = J^{tr} A J$, where J is the Jacobian, i.e.

$$A = \begin{pmatrix} 1 & 0 \\ 0 & x^2 y \end{pmatrix} \quad J = \begin{pmatrix} 2x & 2x \\ \frac{2}{\sqrt{y}} & -\frac{2}{\sqrt{y}} \end{pmatrix}$$

we obtain

$$\tilde{a} = 0 = \tilde{c} \quad \tilde{b} = 8x^2 = 4(\xi + \eta)$$

so that the principal part of the PDE in transformed coordinates is $8(\xi + \eta)u_{\xi\eta}$. We leave it as an exercise that the PDE transforms to

$$u_{\xi\eta} + \frac{3\xi + \eta}{4(\eta^2 - \xi^2)}u_{\xi} - \frac{\xi + 3\eta}{4(\xi^2 - \eta^2)}u_{\eta} = 0 \quad (\xi > \eta)$$

2. **Problem:** Transform the equation $e^{2x}u_{xx} + 2e^{x+y}u_{xy} + e^{2y}u_{yy} = 0$ to a canonical form without mixed partial derivatives.

Solution: We have $b^2 - ac = 0$ so that this PDE is parabolic. We therefore expect to be able to reduce it to something like the heat equation. Specifically, we seek coordinates ξ, η so that the principal part of the PDE is $u_{\xi\xi}$.

First, we determine the characteristics, i.e. the solutions of

$$e^{2x}(dy)^2 - 2e^{x+y}(dx)(dy) + e^{2y}(dx)^2 = 0$$

This yields $(\frac{dy}{dx})^2 - 2e^{y-x}\frac{dy}{dx} + e^{2(y-x)} = 0$, so that

$$\frac{dy}{dx} = e^{y-x} \quad \text{i.e. } e^{-x} dx - e^{-y} dy = 0$$

and hence the characteristics are the curves $e^{-x} - e^{-y} = \text{const.}$. We seek coordinates ξ, η so that the PDE will have principal part $u_{\xi\xi}$, i.e. we want $\tilde{b} = 0 = \tilde{c}$. Now $\tilde{c} = a\eta_x^2 + 2b\eta_x\eta_y + c\eta_y^2$, and thus to ensure that $\tilde{c} = 0$ we see that η must satisfy the characteristic equation. As the PDE remains parabolic under the transformation, we then will also have $\tilde{b}^2 - \tilde{a}\tilde{c} = 0$, which immediately implies that $\tilde{b} = 0$ as well, as desired. We thus define η to be a solution of the characteristic equation, i.e. $\eta = e^{-x} - e^{-y}$. As for ξ , it can be anything (subject to the constraint that $\xi_x\eta_y - \xi_y\eta_x \neq 0$). We chose $\xi = x$. Then $\tilde{a} = a\xi_x^2 + 2b\xi_x\xi_y + c\xi_y^2 = e^{2x}$. Some calculations show that the PDE becomes $e^{2x}u_{\xi\xi} + 2u_{\eta} = 0$, and thus that

$$u_{\xi\xi} + 2e^{-2\xi}u_{\eta} = 0$$

□

Remarks 13.4.5 1. The characteristic curves of a second-order PDE $au_{xx} + 2b_{uxy} + cu_{yy} + du_x + eu_y + fu = g$ can also be interpreted as *exceptional curves*: The PDE together with the values u, u_x, u_y along the curve do not uniquely determine the values of u_{xx}, u_{xy}, u_{yy} . Let us see what is meant by this: Take a smooth curve C in \mathbb{R}^2 , parametrized by a coordinate s , i.e. there is an interval I and maps $x, y : I \rightarrow \mathbb{R}$ so that C is the set of all points $(x(s), y(s))$ for $s \in I$. Suppose that u, u_x, u_y are known along C , i.e. there are functions $F, G, H : I \rightarrow \mathbb{R}$ so that

$$u(x(s), y(s)) = F(s) \quad u_x(x(s), y(s)) = G(s) \quad u_y(x(s), y(s)) = H(s)$$

Then we obtain three equations in the three unknowns u_{xx}, u_{xy}, u_{yy} at the point $(x(s), y(s))$. The first is the PDE, and the other two are obtained by differentiating $G = u_x, H = u_y$

w.r.t. s :

$$au_{xx} + 2bu_{xy} + cu_{yy} = g - dF(s) - eG(s) - fF(s)$$

$$u_{xx}x'(s) + u_{xy}y'(s) = \frac{du_x}{ds}$$

$$u_{xy}x'(s) + u_{yy}y'(s) = H'(s)$$

where a, b, \dots, g are functions of s : $a = a(x(s), y(s))$, etc.

To solve it we look at the coefficient matrix

$$A := \begin{pmatrix} a & 2b & c \\ x'(s) & y'(s) & 0 \\ 0 & x'(s) & y'(s) \end{pmatrix}$$

We will not be able to solve this precisely when $\det(A) = 0$, i.e. when

$$a\left(\frac{dy}{ds}\right)^2 - 2b\frac{dx}{ds}\frac{dy}{ds} + c\left(\frac{dx}{ds}\right)^2 = 0$$

which implies the defining equation of the characteristics

$$a(dy)^2 - 2b(dx)(dy) + c(dx)^2 = 0$$

Thus the characteristics are precisely those curves along which the second derivatives are indeterminate (given the PDE and lower order derivatives).

2. Related to the above is the following fact: the characteristics are the only curves along which *discontinuities* in the solution u and its derivatives can occur. We won't prove this here, but it may be seen intuitively as follows: If u, u_x, u_y are continuous in a region, then u_{xx}, u_{xy}, u_{yy} , being determined by u, u_x, u_y along any smooth curve in that region — except the characteristics! — will be continuous also. So if any of the second-order derivatives is discontinuous, it must be along a characteristic curve.

This has important consequences for the nature of a solution of a PDE:

- Laplace's equation has no characteristic curves; hence solutions *must* be smooth.
- For the heat equation has only one set of characteristic curves, namely curves $t = \text{constant}$. So discontinuities can only occur along curves where t is constant. Since we are interested in the *time evolution* of the solution to the heat equation, we never hold t constant. Hence if discontinuities occur at $t = 0$, they cannot spread into the region $t > 0$. So even if the initial temperature has discontinuities, the temperature will be smooth at any later time.
- The wave equation $u_{tt} - \alpha^2 u_{xx}$ has two sets of characteristics $x \pm \alpha t$, which extend from the initial time $t = 0$ into the region $t > 0$. Hence singularities can propagate, along these characteristic curves. It is for hyperbolic PDEs that characteristic curves are most important, and we will therefore make scant use of them from here on.

□

For n -dimensional second order linear PDEs, the characteristic surfaces are the level surfaces

$$z(x_1, \dots, x_n) = \text{const.}$$

of solutions of the characteristic equation

$$\sum_{i,j} a_{ij} z_{x_i x_j} = 0$$

This cannot usually be reduced to an ODE if $n > 2$, and other techniques need to be used. However, when the PDE has constant coefficients, we can always use a orthogonal change of variables — using an orthonormal basis of eigenvectors — to remove the mixed partial derivatives.

Example 13.4.6 Consider the PDE

$$3u_{xx} - 2u_{xy} + 2u_{yy} - 2u_{yz} + 3u_{zz} + 12u_y - 8u_z = 0$$

The principal part has matrix

$$A = \begin{pmatrix} 3 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 3 \end{pmatrix}$$

The eigenvalues are given by the equation $\det(A - \lambda I) = 0$, and are found to be $\lambda = 1, 3, 4$. All the eigenvalues are strictly positive, so this PDE is elliptic. Corresponding orthonormal eigenvectors are easily determined to be $\frac{1}{\sqrt{6}}(1, 2, 1)^{tr}$, $\frac{1}{\sqrt{2}}(1, 0, -1)^{tr}$, $\frac{1}{\sqrt{3}}(1, -1, 1)^{tr}$. Thus we use the transformation $\xi = O^{tr}\mathbf{x}$, where O is the orthogonal matrix whose columns are the orthonormal eigenvectors of A , i.e.

$$\begin{pmatrix} \xi \\ \eta \\ \chi \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

to obtain a PDE with principal part

$$u_{\xi\xi} + 3u_{\eta\eta} + 4u_{\chi\chi}$$

□

Chapter 14

Laplace's Equation

14.1 The Divergence Theorem and Related Results

This section serves to remind you of one of the basic tools in PDE theory: The Divergence (or Gauss–Green) Theorem. Let $U \subseteq \mathbb{R}^m$ be a bounded open set, with a smooth boundary ∂U , and let \mathbf{n} denote the outward pointing unit normal vector to the boundary ∂U . (\mathbf{n} is actually a *vector field*, i.e. a map $\mathbf{n} : \partial U \rightarrow \mathbb{R}^m$: We have an outward pointing normal vector $\mathbf{n}_{\mathbf{x}}$ at every point $\mathbf{x} \in \partial U$.) For any $u \in C^1(\bar{U})$, we denote the directional derivative in the direction of \mathbf{n} by $\frac{\partial u}{\partial n}$:

$$\frac{\partial u}{\partial n} \Big|_{\mathbf{x}} = Du(\mathbf{x}) \cdot \mathbf{n}_{\mathbf{x}}$$

Lemma 14.1.1 *If U is a ball centered at the origin, and r is the radial component of (n -dimensional) polar coordinates then*

$$\frac{\partial}{\partial n} = \frac{\partial}{\partial r}$$

Proof: For the unit normal at point \mathbf{x} is $\mathbf{n} = \frac{\mathbf{x}}{\|\mathbf{x}\|} = \frac{\mathbf{x}}{r}$, where $r = (\sum_{k=1}^n x_k^2)^{\frac{1}{2}}$. Note also that $\frac{\partial x_k}{\partial r} = \frac{x_k}{r}$, because each coordinate x_k has polar coordinates of the form $x_k = r \prod_j \text{trig}_j(\theta_j)$, where each function trig_j is either sin or cos, and θ_j are angles. Hence

$$\frac{\partial u}{\partial n} \Big|_{\mathbf{x}} = Du(\mathbf{x}) \cdot \mathbf{n} = \sum_{k=1}^n u_{x_k} \frac{x_k}{r} = \sum_{k=1}^n \frac{\partial u}{\partial x_k} \frac{\partial x_k}{\partial r} = \frac{\partial u}{\partial r} \Big|_{\mathbf{x}}$$

+

All forms of the Divergence Theorem follow from the following basic tool:

Proposition 14.1.2 *Suppose that $u \in C^1(\bar{U})$. Then*

$$\int_U u_{x_i} d\mathbf{x} = \oint_{\partial U} u n_i dS$$

(where $u_{x_i} := \frac{\partial u}{\partial x_i}$ and $\mathbf{n} = (n_1, \dots, n_m)$ is the outward pointing unit normal.)

Proof: We give an outline of the proof in three dimensions, with $x_i = z$ -coordinate. The same ideas work in higher dimensions (but note that the proof makes heavy use of spatial intuition, and that a rigorous proof is actually quite difficult, even in three dimensions.) A *simple* region is a region which is cylindrical in the direction of one of the coordinate axes, with *smooth* disjoint top and bottom surfaces. A sufficiently regular bounded open set U can be decomposed into many “cubes” C_i , which will be simple. The volume integrals $\int_{C_i} u_z dV$ over each component cube will add up to the volume integral over the whole of U : $\sum_i \int_{C_i} f_z d\mathbf{x} = \int_U u_z d\mathbf{x}$. If two cubes are adjacent, i.e. share a common surface, then the surface integrals over that common surface will cancel each other out, as the outward unit normal of the one will be minus the outward unit normal of the other. Thus the surface integrals of all the cubes will add up to the surface integral over ∂U : $\sum_i \oint_{\partial C_i} u_n dS = \oint_{\partial U} u_n dS$. It follows that if the divergence theorem holds for each of these simple cubes, i.e. if $\int_{C_i} u_z d\mathbf{x} = \oint_{\partial C_i} u_n dS$ for all i , then it will hold for the whole region U :

$$\int_U u_z d\mathbf{x} = \sum_i \int_{C_i} u_z d\mathbf{x} = \sum_i \oint_{\partial C_i} u_n dS = \oint_{\partial U} u_n dS$$

Consider now a simple “cube” C , cylindrical in the direction of the z -axis, with smooth top and bottom. Let the top surface be given by $z = t(x, y)$ and the bottom surface by $z = b(x, y)$, where b, t are C^1 -functions. Also let A be the projection of the cube onto the xy -plane, so that A is a rectangle in the xy -plane, with sides parallel to the axes. Thus

$$C = \{(x, y, z) \in \mathbb{R}^3 : (x, y) \in A \text{ and } b(x, y) \leq z \leq t(x, y)\}$$

We first calculate $\oint_{\partial C} u_n dS$. On the lateral sides, the z -component of the outward unit normal is zero, so the surface integrals over the lateral surfaces do not contribute to $\oint_{\partial C} u_n dS$. We thus have

$$\oint_{\partial C} u_n dS = \int_{\text{top}} u_n dS + \int_{\text{bottom}} u_n dS$$

Let's look at the integral over the top surface: If θ is the (acute) angle between the outward unit normal of the top surface and the z -axis, then $dS = \frac{dA}{\cos \theta} = \frac{dx dy}{\cos \theta}$. Moreover, $n_z = \cos \theta$. Hence $\int_{\text{top}} u(x, y, z) n_z dS = \int_A u(x, y, t(x, y)); dx dy$.

Now consider the integral over the bottom surface: If φ is the (obtuse) angle between the outward unit normal of the bottom surface and the z -axis, then $\cos \varphi \leq 0$. It follows that $dS = -\frac{dx dy}{\cos \varphi}$ and $n_z = \cos \varphi$, so that $\int_{\text{bottom}} u(x, y, z) n_z dS = -\int_A u(x, y, b(x, y)); dx dy$. It follows that

$$\int_{\partial C} u_n dS = \int_A u(x, y, t(x, y)) - u(x, y, b(x, y)) dx dy$$

But

$$\int_C u_z d\mathbf{x} = \int_A \left(\int_{b(x, y)}^{t(x, y)} u_z dz \right) dx dy = \int_A u(x, y, t(x, y)) - u(x, y, b(x, y)) dx dy$$

by the Fundamental Theorem of Calculus, because $\int_b^t \frac{\partial}{\partial z} u(x, y, z) dz = u(x, y, t) - u(x, y, b)$. Hence $\int_C u_z d\mathbf{x} = \oint_{\partial C} u_n dS$, as required, and the result follows.

+

We have the following corollaries:

Theorem 14.1.3 (Integration-by-parts formula) *Suppose that $u, v \in C^1(\bar{U})$, where $U \subseteq \mathbb{R}^m$ is a bounded open set with smooth boundary. Then*

$$\int_U u_{x_i} v \, d\mathbf{x} = - \int_U uv_{x_i} \, d\mathbf{x} + \oint_{\partial U} uv \, n_i \, dS$$

□

Theorem 14.1.4 (Divergence Theorem: vector form) *If $\mathbf{u} : \mathbb{R}^m \rightarrow \mathbb{R}^m$ is continuously differentiable, where $U \subseteq \mathbb{R}^m$ is a bounded open set with smooth boundary. Then*

$$\int_U \operatorname{div} \mathbf{u} \, d\mathbf{x} = \oint_{\partial U} \mathbf{u} \cdot \mathbf{n} \, dS$$

□

Theorem 14.1.5 (Green's Formulas) *Suppose that $u, v \in C^2(\bar{U})$, where $U \subseteq \mathbb{R}^m$ is a bounded open set with smooth boundary. Then*

- I. $\int_U \Delta u \, dV = \oint_{\partial U} \frac{\partial u}{\partial n} \, dS$
- II. $\int_U Du \cdot Dv \, d\mathbf{x} = - \int_U u \Delta v \, d\mathbf{x} + \oint_{\partial U} \frac{\partial v}{\partial n} u \, dS$
- III. $\int_U u \Delta v - v \Delta u \, d\mathbf{x} = \oint_{\partial U} u \frac{\partial v}{\partial n} - v \frac{\partial u}{\partial n} \, dS$

□

Exercise 14.1.6 Prove the preceding theorems.

[**Hints:** For the integration by parts formula, apply Proposition 14.1.2 to the function uv . For the vector form of the divergence theorem, apply Proposition 14.1.2 to each of the components of the vector-valued function \mathbf{u} and sum.

For Green I, use the integration-by-parts formula with u_{x_i} in place of u and $v = 1$.

For Green II, use integration-by-parts with v_{x_i} in place of v .

For Green III, use Green II.]

□

14.2 Harmonic Functions

In this section, we study *harmonic* functions. These are exactly the solutions to Laplace's equation, i.e. a C^2 -function $u : \mathbb{R}^n \rightarrow \mathbb{R}$ is said to be harmonic iff

$$\sum_{i=1}^n \frac{\partial^2 u}{\partial x_i^2} = 0 \quad \text{i.e.} \quad \Delta u = 0$$

14.2.1 Some Heuristic Remarks about the Laplace Operator

The Laplace operator in \mathbb{R}^n is

$$\Delta := \sum_{i=1}^n \frac{\partial^2}{\partial x_i^2}$$

In one-dimension, to say that $\Delta u(x) \geq 0$ implies that if and only if u is concave up at x . This implies that if we take two neighbouring points $x - a, x + a$, we have

$$u(x) \leq \frac{1}{2}u(x - a) + \frac{1}{2}u(x + a)$$

i.e. the value $u(x)$ is smaller than the average of its neighbours. In higher dimensions, $\Delta u(\mathbf{x}) \geq 0$ does not imply that the function is concave up: Consider, for example, $u(x, y) := x^4 - y^2$ which is concave up along the x -axis (where $y = 0$) and concave down along the y -axis (where $x = 0$). (Indeed, we know that the correct generalization of the second derivative test is that the Hessian matrix be positive definite. Here, the Hessian is $\begin{pmatrix} 12 & 0 \\ 0 & -2 \end{pmatrix}$ which is not positive definite.) However, the averaging property does hold in higher dimensions:

If $\Delta u \geq 0$, then the value of u at \mathbf{x} is smaller than the average the values of its neighbours.

This is an imprecise statement, whose purpose is only to provide intuition. We shall formalize it in the next section, when we discuss the mean value property. For the moment, however, let's give an intuitive argument in two dimensions — the argument can easily be generalized to higher dimensions. Note that a second-order Taylor expansion yields

$$u(\mathbf{x} + \Delta \mathbf{x}) = u(\mathbf{x}) + Du(\mathbf{x}) \cdot \Delta \mathbf{x} + \frac{1}{2} \Delta \mathbf{x}^{tr} Hu(\mathbf{x}) \Delta \mathbf{x} + o(\|\Delta \mathbf{x}\|^2)$$

where $Hu(\mathbf{x}) = \begin{pmatrix} u_{xx} & u_{xy} \\ u_{yx} & u_{yy} \end{pmatrix} \big|_{\mathbf{x}}$ is the Hessian matrix of u evaluated at the point \mathbf{x} . Fix $h \in \mathbb{R}$, substitute the following four values of $\Delta \mathbf{x}$ into the Taylor expansion

$$\Delta \mathbf{x} = (h, h) \quad \Delta \mathbf{x} = (h, -h) \quad \Delta \mathbf{x} = (-h, 0) \quad \Delta \mathbf{x} = (-h, -h)$$

and then add to obtain

$$\begin{aligned} & \frac{u(x+h, y+h) + u(x+h, y-h) + u(x-h, y+h) + u(x-h, y-h)}{4} \\ &= u(x, y) + \frac{1}{2}(u_{xx} + u_{yy})h^2 + o(h^2) \geq u(x, y) \end{aligned}$$

We thus see that $u(x, y)$ is less than the average of values at the neighbouring points $(x \pm h, y \pm h)$.

As stated before, we will make this precise in the next section. For the moment, note that the three canonical examples of second order linear PDEs are

$$\begin{aligned} 0 &= \Delta u && \text{Laplace's equation} \\ u_t &= \Delta u && \text{heat equation} \\ u_{tt} &= \Delta u && \text{wave equation} \end{aligned}$$

We can now make some qualitative remarks about the solutions of these PDEs:

- If u satisfies Laplace's equation, it means that the value of u at a point \mathbf{x} equals the average of the values of nearby points.
- If u satisfies the heat equation, then $u_t \geq 0$ when $\Delta u \geq 0$. This means that the temperature at a point \mathbf{x} will increase ($u_t \geq 0$) when it is less than the average temperature at neighbouring points.
- If u satisfies the wave equation, then $u_{tt} \geq 0$ when $\Delta u \geq 0$. This means that a point at position \mathbf{x} on a vibrating string will accelerate upwards when its displacement is less than the average displacement at neighbouring points.

14.2.2 Volumes and Surface Areas of Balls in \mathbb{R}^n

We will leave the proof of the following theorem as an exercise, as we shall only apply it in the case $n \leq 3$, where the result is presumably well-known:

Theorem 14.2.1 *In \mathbb{R}^n , the “volume” $V_n(R)$ and “surface area” $A_n(R)$ of a ball with radius R is given by*

$$A_n(R) = \frac{2\pi^{n/2}}{\Gamma(n/2)} = \begin{cases} \frac{n\pi^{n/2}}{(n/2)!} R^{n-1} & \text{if } n \text{ is even} \\ \frac{2n(2\pi)^{(n-1)/2}}{1 \cdot 3 \cdot 5 \cdots n} R^{n-1} & \text{if } n \text{ is odd} \end{cases} \quad V_n(R) = A_n(R) \frac{R}{n}$$

(where $\Gamma(x) := \int_0^\infty e^{-t} t^{x-1} dt$ is the Euler Gamma function).

□

Exercise 14.2.2 We prove the above result.

- (a) First, we show that $\int_{\mathbb{R}^n} e^{-\|\mathbf{x}\|^2} d\mathbf{x} = \pi^{n/2}$. To prove this, note that if $C := \int_{-\infty}^\infty e^{-x^2} dx$, then

$$C^n = \prod_{j=1}^n \left(\int_{-\infty}^\infty e^{-x_j^2} dx_j \right) = \int_{\mathbb{R}^n} e^{-\|\mathbf{x}\|^2} d\mathbf{x}$$

For $n = 2$, use polar coordinates to deduce that

$$C^2 = \int_0^{2\pi} \int_0^\infty e^{-r^2} r dr d\theta = \pi$$

so that $C = \pi^{1/2}$.

- (b) Converting to n -dimensional polar coordinates, note that

$$\int_{\mathbb{R}^n} e^{-\|\mathbf{x}\|^2} d\mathbf{x} = A_n(1) \int_0^\infty e^{-r^2} r^{n-1} dr$$

where $A_n(r)$ denotes the surface area of an n -dimensional ball of radius r . This formula will also hold for $n = 1$, provided we define $A_1(1) = 2$.

- (c) Conclude that

$$\pi^{n/2} = \frac{1}{2} A_n(1) \Gamma\left(\frac{n}{2}\right)$$

(d) Show that the Gamma function Γ has the following properties:

- (i) $\Gamma(x+1) = x\Gamma(x)$.
- (ii) $\Gamma(0) = 1$. Hence $\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$.
- (iii) $\Gamma(\frac{1}{2}) = \pi^{\frac{1}{2}}$. Hence $\Gamma(n + \frac{1}{2}) = \pi^{\frac{1}{2}}(n - \frac{1}{2})(n - \frac{3}{2}) \dots (\frac{1}{2})$.

[Hint: For (i), use integration by parts. For (iii), use (c) and the fact that $A_1(1) = 2$]

(e) Now verify that the formulas for areas $A_n(r)$ given in Thm. 14.2.1 are correct.

(f) Finally, show that

$$V_n(R) = \int_0^R A_n(1)r^{n-1} dr = A_n(R)\frac{R}{n}$$

□

14.2.3 Mean-Value Property and the Maximum Principle

Definition 14.2.3 A function u defined on an open set $U \subseteq \mathbb{R}^n$ is said to satisfy the *mean-value property* at $\mathbf{x}_0 \in U$ if

$$u(\mathbf{x}_0) = \frac{1}{A_n(R)} \oint_{\partial B(\mathbf{x}_0, R)} u(\mathbf{x}) dS_R$$

for every ball $B(\mathbf{x}_0, R)$ of radius $R > 0$ centered at \mathbf{x}_0 which is contained in U .

□

The integral in the definition above is a surface integral over the surface (i.e. boundary) of the ball $B(\mathbf{x}_0, R)$.

Theorem 14.2.4 (Mean Value Property of Harmonic Functions) *Let $U \subseteq \mathbb{R}^n$ be an open set.*

- (a) *If $u \in C^2(U) \cap C^0(\bar{U})$ is harmonic in U , then u satisfies the mean value property at each $\mathbf{x}_0 \in U$.*
- (b) *Conversely, if function $u \in C^2(U) \cap C^0(\bar{U})$ satisfies the mean value property, then it is harmonic.*

Proof: (a) For $r > 0$, define

$$\phi(r) := \frac{1}{A_n(r)} \oint_{\partial B(\mathbf{x}_0, r)} u(\mathbf{x}) dS = \frac{1}{A_n(1)} \oint_{\partial B(0,1)} u(\mathbf{x}_0 + r\mathbf{z}) dS(\mathbf{z})$$

Then

$$\phi'(r) = \frac{1}{A_n(1)} \oint_{\partial B(0,1)} Du(\mathbf{x}_0 + r\mathbf{z}) \cdot \mathbf{z} dS(\mathbf{z}) = \frac{1}{A_n(r)} \oint_{\partial B(\mathbf{x}_0, r)} Du(\mathbf{x}_0) \cdot \frac{\mathbf{x} - \mathbf{x}_0}{r} dS(\mathbf{x})$$

Now $\mathbf{n} := \frac{\mathbf{x} - \mathbf{x}_0}{r}$ is precisely the outward unit normal to $\partial B(\mathbf{x}_0, r)$ at \mathbf{x} , so by the divergence theorem (noting that $\operatorname{div} \cdot Du = \Delta u$) we get

$$\phi'(r) = \frac{1}{A_n(r)} \int_{B(\mathbf{x}_0, r)} \Delta u(\mathbf{x}) \, d\mathbf{x}$$

Since u is harmonic on U , $\Delta u = 0$ on $B(\mathbf{x}_0, r) \subseteq U$, and so $\phi'(r) = 0$. It follows that ϕ is constant. In particular

$$\frac{1}{A_n(r)} \oint_{\partial B(\mathbf{x}_0, r)} u(\mathbf{x}) \, dS(\mathbf{x}) = \phi(r) = \lim_{s \rightarrow 0} \phi(s) = \lim_{s \rightarrow 0} \oint_{\partial B(\mathbf{x}_0, s)} u(\mathbf{x}) \, dS(\mathbf{x}) = u(\mathbf{x}_0)$$

which proves (a).

(b) Suppose that u has the mean value property, but is not harmonic. Then there is an open set $V \subseteq U$ where $\Delta u > 0$, or else one where $\Delta u < 0$. Without loss of generality, assume the former holds, i.e. that $\Delta u > 0$ on $V \subseteq U$. Let $\mathbf{x}_0 \in V$, and define $\phi(r)$ as in (a). Note that we obtained there the following equation:

$$\phi'(r) = \frac{1}{A_n(r)} \int_{B(\mathbf{x}_0, r)} \Delta u(\mathbf{x}) \, d\mathbf{x}$$

It follows that $\phi'(r) > 0$. On the other hand, the mean value property at \mathbf{x}_0 is precisely the statement that $\phi(r)$ is a constant function with value $u(\mathbf{x}_0)$ — contradiction.

—

As an easy corollary of (a), we prove a “solid” version of the mean value property:

Corollary 14.2.5

If a twice continuously differentiable function u is harmonic in an open set $U \subseteq \mathbb{R}^n$, then

$$u(\mathbf{x}_0) = \frac{1}{V_n(r)} \int_{B(\mathbf{x}_0, r)} u(\mathbf{x}) \, d\mathbf{x}$$

for every $r > 0$ such that $B(\mathbf{x}_0, r) \subseteq U$.

Proof: Clearly

$$\begin{aligned} \frac{1}{V_n(r)} \int_{B(\mathbf{x}_0, r)} u(\mathbf{x}) \, d\mathbf{x} &= \frac{1}{V_n(r)} \int_0^r \left(\oint_{\partial B(\mathbf{x}_0, s)} u(\mathbf{x}) \, dS(\mathbf{x}) \right) ds \\ &= \frac{u(\mathbf{x}_0)}{V_n(r)} \int_0^r A_n(s) \, ds \\ &= u(\mathbf{x}_0) \end{aligned}$$

—

An intuitively plausible consequence of the mean value property is that harmonic functions are very smooth: If u is harmonic, then the value $u(x)$ of u at x is the average of the $u(y)$ for neighbouring points y , and this averaging ensures smoothness. We won't show it here, but it can be proved that harmonic functions are *analytic*, i.e. C^∞ -functions (infinitely

differentiable) which are everywhere equal to their Taylor series expansions (in their domain of convergence).

Another important consequence of the mean value property is the following, the *strong maximum principle*: The maximum of a harmonic function on a bounded open set occurs on the boundary. Furthermore if an interior point is a maximum, then u is constant:

Theorem 14.2.6 (a) *Let U be a bounded open set, and suppose that u is harmonic in U and continuous on \bar{U} . Then the maximum of*

$$\max_{\mathbf{x} \in \bar{U}} u(\mathbf{x}) = \max_{\mathbf{x} \in \partial U} u(\mathbf{x})$$

i.e. the maximum occurs on the boundary.

(b) *Furthermore, if U is connected and there is an interior point $\mathbf{x}_0 \in U$ such that*

$$u(\mathbf{x}_0) = \max_{\mathbf{x} \in \bar{U}} u(\mathbf{x})$$

then u is constant on U .

Proof: We first prove (b): If the maximum of u on \bar{U} occurs at some interior point $\mathbf{x}_0 \in U$, then since

$$u(\mathbf{x}_0) = \frac{1}{V_n(r)} \int_{B(\mathbf{x}_0, r)} u(\mathbf{x}) \, d\mathbf{x}$$

we see that we must have $u(\mathbf{x}) = u(\mathbf{x}_0)$ for every $\mathbf{x} \in B(\mathbf{x}_0, r)$. Now given an arbitrary point $\mathbf{y} \in U$ we can connect \mathbf{x}_0 to \mathbf{y} by a finite sequence of overlapping open balls in U (i.e. balls $B_0, B_1, \dots, B_m \subseteq U$ so that \mathbf{x}_0 is the center of B_0 , \mathbf{y} the center of B_m and $B_{i-1} \cap B_i \neq \emptyset$ for $i = 1, \dots, m$.) Then u is constant on each ball, and hence $u(\mathbf{x}_0) = u(\mathbf{y})$. Since $\mathbf{y} \in U$ was arbitrary, u is constant on U . This proves (b).

(a) follows straightaway: If u is constant, then of course the maximum occurs on the boundary (and everywhere else also). If u is not constant, then the maximum cannot occur at an interior point, by (b), so must occur at a boundary point.

—

Now when u is harmonic, so is $-u$. Applying the maximum principle to $-u$ leads to an analogous “minimum principle” for u : The minimum of a harmonic function on a bounded open set occurs on the boundary. Moreover, if it also occurs in the interior, then u is constant.

Exercise 14.2.7 Interpret and verify the results of this section for one-dimensional harmonic functions, i.e. functions $u(x)$ having $u'' = 0$.

□

Exercise 14.2.8 Laplace's equation is invariant under translations and rotations: If $u : \mathbb{R}^n \rightarrow \mathbb{R}$ has $\Delta u = 0$, and if

$$v(\mathbf{x}) := u(\mathbf{x} + \mathbf{c}) \quad w(\mathbf{x}) := u(O\mathbf{x})$$

where $\mathbf{c} \in \mathbb{R}^n$, and O is an orthogonal $n \times n$ -matrix, then $\Delta v = 0$ and $\Delta w = 0$.

[Recall from linear algebra that a matrix is *orthogonal* iff $O^{-1} = O^{tr}$. The orthogonal matrices are precisely the linear transformations which are rotations: For if $\mathbf{x} \in \mathbb{R}^n$, then

$$\|O\mathbf{x}\|^2 = (O\mathbf{x})^{tr} \cdot (O\mathbf{x}) = \mathbf{x}^{tr} O^{tr} O \mathbf{x} = \mathbf{x}^{tr} \mathbf{x} = \|\mathbf{x}\|^2$$

i.e. $\|O\mathbf{x}\| = \|\mathbf{x}\|$.]

□

Exercise 14.2.9 Here is a proof of the *maximum principle*. It is weaker than the proof based on the mean value property, as it doesn't show the absence of maxima at interior points, but only that there are boundary maxima. Nevertheless, it is quite intuitive, appealing only to the second derivative test for maxima, and will improve understanding.

Let $u \in C^2(U) \cap C^0(\bar{U})$ be *subharmonic* on U , where $U \subseteq \mathbb{R}^n$ is a bounded open set, i.e. assume that $\Delta u \geq 0$. Since U is bounded, there is R such that $\bar{U} \subseteq B(0, R)$.

- (a) Suppose that u has a maximum at an interior point $\mathbf{x}_0 \in U$. Explain why $u_{x_i x_i}(\mathbf{x}_0) \leq 0$ for all $i = 1, \dots, n$.
- (b) Conclude that $\Delta u \leq 0$.
- (c) At *most* maxima, we have $u_{x_i x_i} < 0$. If this is the case, we have $\Delta u < 0$, contradicting the fact that u is subharmonic. But it is possible that $\Delta u = 0$ at a minimum. Give an example of such a u .
- (d) To get around this, define

$$v(\mathbf{x}) := u(\mathbf{x}) + \varepsilon \|\mathbf{x}\|^2$$

for $\varepsilon > 0$. Show that

$$\Delta v = \Delta u + 2n\varepsilon > 0$$

where n is the dimension of the space.

- (e) Conclude that v has no maximum in the interior U .
- (f) Explain why v must have a maximum on \bar{U} , and conclude that v has a maximum at some point $\mathbf{x}_0 \in \partial U$ on the boundary.
- (g) Conclude that if $\mathbf{x} \in U$ is in the interior, then

$$u(\mathbf{x}) < v(\mathbf{x}) \leq v(\mathbf{x}_0) \leq u(\mathbf{x}_0) + \varepsilon \|\mathbf{x}_0\|^2 \leq \sup_{\mathbf{y} \in \partial U} u(\mathbf{y}) + \varepsilon R^2$$

- (h) Explain why we may conclude that for all interior points $\mathbf{x} \in U$ we have

$$u(\mathbf{x}) \leq \sup_{\mathbf{y} \in \partial U} u(\mathbf{y})$$

- (i) Also explain why there is $\mathbf{x}_M \in \partial U$ such that $\sup_{\mathbf{y} \in \partial U} u(\mathbf{y}) = u(\mathbf{x}_M)$.
- (j) Conclude that a maximum of u occurs at a boundary point.

□

Remarks 14.2.10 Here are just a few comments for those who know some complex analysis. A holomorphic (i.e. differentiable) function $f : \mathbb{C} \rightarrow \mathbb{C}$ may be regarded as a pair of real-valued functions $u, v : \mathbb{R}^2 \rightarrow \mathbb{R}$:

$$f(z) = f(x + iy) = u(x, y) + iv(x, y)$$

But if a function f is holomorphic, then the Cauchy–Riemann equations must hold:

$$u_x = v_y \quad \text{and} \quad u_y = -v_x$$

(Recall that these are obtained by computing $f'(z)$ in two ways: On the one hand $f'(z) = \lim_{\Delta x \rightarrow 0} \frac{f((x+\Delta x)+iy) - f(x+iy)}{\Delta x} = u_x(x, y) + iv_x(x, y)$. But on the other hand, we have $f'(z) = \lim_{\Delta y \rightarrow 0} \frac{f(x+i(y+\Delta y)) - f(x+iy)}{i\Delta y} = -iu_y(x, y) + v_y(x, y)$. Hence $u_x + iv_x = f' = v_y - iu_y$. Equating real and imaginary parts, one obtains the Cauchy–Riemann equations.) Differentiating the first of the Cauchy–Riemann equations with respect to x , and the second with respect to y , we see that $u_{xx} = v_{yx}$ and $u_{yy} = -v_{xy}$. As $v_{xy} = v_{yx}$ we see that

$$u_{xx} + u_{yy} = 0$$

In the same manner, differentiating the first equation w.r.t. y and the second w.r.t. x one derives,

$$v_{xx} + v_{yy} = 0$$

Hence the real and imaginary parts of a holomorphic function are harmonic in \mathbb{R}^2 .

In complex analysis, the maximum modulus theorem is the analogue of the maximum principle. The mean value property, which states that a harmonic function is completely determined by its values on the boundary of a curve, is clearly related to the uniqueness of *analytic continuation*. It is also well-known that holomorphic functions are *analytic*, i.e. C^∞ and everywhere equal to their Taylor series expansion, and a similar result can be proved for n -dimensional harmonic functions.

□

14.3 Solving Laplace's Equation

14.3.1 Uniqueness of Solutions

The canonical example of an elliptic equation is Laplace's equation

$$\Delta u = 0$$

and its non-homogeneous version, Poisson's equation

$$-\Delta u = f$$

The solutions of Laplace's equation are precisely the harmonic functions. Problems which are well-posed for Poisson's or Laplace's equation will usually be well-posed for more general elliptic problems as well. The maximum principle allows us to immediately prove a partial well-posedness result for the inhomogeneous Dirichlet problem for the Poisson equation:

Theorem 14.3.1 *Let $U \subseteq \mathbb{R}^n$ be a bounded open set. The Dirichlet problem*

$$\begin{aligned}\Delta u &= f && \text{in } U \\ u &= g && \text{on } \partial U\end{aligned}$$

has at most one solution.

Proof: If u_1, u_2 are two solutions to the above Dirichlet problem, then $u := u_1 - u_2$ satisfies the corresponding homogeneous problem

$$\begin{aligned}\Delta u &= 0 && \text{in } U \\ u &= 0 && \text{on } \partial U\end{aligned}$$

Thus u is harmonic. Since its maximum and minimum must occur on the boundary, where u takes the value 0, we see that $0 \leq u(\mathbf{x}) \leq 0$ for all $\mathbf{x} \in \bar{U}$. It follows that $u_1 = u_2$, i.e. that the solution, if it exists, is unique.

—

Remarks 14.3.2 Note the contrast with the wave equation: Consider a vibrating string of length 1, fixed at its endpoints. Let $u(t, x)$ be the displacement at time t at point x , where $0 \leq x \leq 1$, and suppose that we seek a solution with the property that the initial displacement is zero, i.e. $u(0, x) = 0$ for all x , and that the displacement at time $T = 1$ is zero, i.e. $u(1, x) = 0$ for all x . Here, we are considering $U = (0, 1) \times (0, 1) \subseteq \mathbb{R}^2$, and the Dirichlet problem to solve is

$$\begin{aligned}u_{tt} - u_{xx} &= 0 && \text{in } U \\ u &= 0 && \text{on } \partial U\end{aligned}$$

For each $n \in \mathbb{N}$, the function

$$u_n(t, x) = \sin(n\pi x) \sin(n\pi t)$$

is clearly a solution, as

$$\frac{\partial^2 u}{\partial t^2} = -n^2 \pi^2 u = \frac{\partial^2 u}{\partial x^2} \quad \text{and} \quad \sin n\pi = 0 = \sin 0$$

Thus comparing the two seemingly similar Dirichlet problems

$$\begin{array}{ll} u_{tt} + u_{xx} = 0 \text{ in } U & \text{(Laplace)} \\ u = 0 \text{ on } \partial U & \end{array} \qquad \begin{array}{ll} u_{tt} - u_{xx} = 0 \text{ in } U & \text{(Wave)} \\ u = 0 \text{ on } \partial U & \end{array}$$

we see that Laplace's equation has a unique solution (namely $u \equiv 0$), whereas the wave equation has many.

As the uniqueness of the solution to the Laplace equation follows from the maximum principle, there can be no analogous maximum principle for the wave equation.

□

14.3.2 Fundamental Solution of Laplace's Equation

Exercise 14.2.8 shows that Laplace's equation is invariant under rotations. This suggests that we seek a *radial* solution to Laplace's equation, i.e. a solution

$$u(\mathbf{x}) = v(r) \quad r = (x_1^2 + x_2^2 + \cdots + x_n^2)^{\frac{1}{2}}$$

which is a function solely of r . Note that

$$\frac{\partial r}{\partial x_i} = \frac{x_i}{r} \quad r \neq 0$$

and so for $i = 1, \dots, n$ we have

$$u_{x_i} = v'(r) \frac{x_i}{r} \quad u_{x_i x_i} = v''(r) \frac{x_i^2}{r^2} + v'(r) \left(\frac{1}{r} - \frac{x_i^2}{r^3} \right)$$

It follows that

$$\Delta u = v''(r) + \frac{n-1}{r} v'(r)$$

and so

$$\Delta u = 0 \quad \Longleftrightarrow \quad v'' + \frac{n-1}{r} v' = 0$$

We thus have a second order ODE to solve. The solution depends on the dimension n : Assuming v is not constant, i.e. $v' \neq 0$, we get

$$\ln(|v'|)' = \frac{v''}{v'} = \frac{1-n}{r}$$

from which $\ln |v'| = (1-n) \ln r + C$, i.e.

$$v'(r) = \frac{A}{r^{n-1}}$$

Integrating again, we see that

$$v(r) = \begin{cases} B \ln r + C & \text{if } n = 2 \\ \frac{B}{r^{n-2}} + C & \text{if } n \geq 3 \end{cases}$$

where B, C are constant.

With the above in mind, we now define

Definition 14.3.3 The function

$$\Psi(\mathbf{x}) := \begin{cases} -\frac{1}{2\pi} \ln \|\mathbf{x}\| & \text{if } n = 2 \\ \frac{1}{(n-2)A_n(1)} \frac{1}{\|\mathbf{x}\|^{n-2}} & \text{if } n \geq 3 \end{cases}$$

defined for $\mathbf{x} \in \mathbb{R}^n - \{0\}$ is called the *fundamental solution* of Laplace's equation. Here, $A_n(1)$ is the surface area of a unit ball in \mathbb{R}^n .

We may on occasion write $\Psi(\|\mathbf{x}\|)$ or $\Psi(r)$ for $\Psi(\mathbf{x})$, to highlight the fact that the fundamental solution Ψ is radially symmetric.

□

Note that Ψ blows up near the origin, and is undefined at the origin. Hence Ψ is harmonic in any region that does not contain the origin.

The reason for the choice of the constants $-\frac{1}{2\pi}$ and $\frac{1}{(n-2)A_n(1)}$ are made clear by the next lemma, which will prove useful a number of times:

Lemma 14.3.4 *If B is an ball of radius R in \mathbb{R}^n , centered at \mathbf{x}_0 , then the outward normal derivative satisfies*

$$\frac{\partial}{\partial n}\Psi(\mathbf{x} - \mathbf{x}_0) = -\frac{1}{A_n(R)}$$

Proof: Let $r = \|\mathbf{x} - \mathbf{x}_0\|$. For $n \geq 3$, we have $\Psi(\mathbf{x} - \mathbf{x}_0) = \frac{1}{(n-2)A_n(1)} \frac{1}{r^{n-2}}$, and for $n = 2$ we have $\Psi(\mathbf{x} - \mathbf{x}_0) = -\frac{1}{2\pi} \ln r$. The outward unit normal derivative on ∂B is given by

$$\frac{\partial}{\partial n} = \frac{\partial}{\partial r}$$

by the same argument as given in the proof of Lemma 14.1.1. But clearly

$$\frac{\partial}{\partial r}\Big|_{r=R} \Psi(r) = \frac{1}{(n-2)A_n(1)} \frac{\partial}{\partial r}\Big|_{r=R} \frac{1}{r^{n-2}} = -\frac{1}{A_n(1)R^{n-1}} = -\frac{1}{A_n(R)}$$

The same argument holds when $n = 2$.

⊥

We can represent all harmonic functions as surface integrals involving the fundamental solution:

Theorem 14.3.5 *Suppose that u is defined on a bounded open set $U \subseteq \mathbb{R}^n$. Then*

$$u(\mathbf{x}_0) = - \int_U \Psi(\mathbf{x} - \mathbf{x}_0) \Delta u(\mathbf{x}) \, d\mathbf{x} - \oint_{\partial U} u(\mathbf{x}) \frac{\partial}{\partial n} \Psi(\mathbf{x} - \mathbf{x}_0) \, dS + \oint_{\partial U} \Psi(\mathbf{x} - \mathbf{x}_0) \frac{\partial u}{\partial n} \, dS$$

for all $\mathbf{x}_0 \in U$.

Proof: Fix $\mathbf{x}_0 \in \mathbb{R}^n$, and define $v(\mathbf{x}) := \Psi(\mathbf{x} - \mathbf{x}_0)$. Note that $\Delta v = 0$. We start with Green's formula III, which states that

$$\int_U u \Delta v - v \Delta u \, d\mathbf{x} = \oint_{\partial U} u \frac{\partial v}{\partial n} - v \frac{\partial u}{\partial n} \, dS \quad (*)$$

But Green's formula $(*)$ holds only when u, v are everywhere defined in U , which is *not* the case here, as $v(\mathbf{x}) = \Psi(\mathbf{x}_0 - \mathbf{x})$ is undefined at $\mathbf{x} = \mathbf{x}_0$.

We therefore *isolate* the singularity \mathbf{x}_0 : Let B_ε be a closed ball of radius ε about \mathbf{x}_0 , contained entirely within U , and let $U_\varepsilon := U - B_\varepsilon$. then U_ε is a bounded open set which does not contain \mathbf{x}_0 , and so u, v are *everywhere* defined on U_ε . Applying Green III, we see that

$$- \int_{U_\varepsilon} v \Delta u \, d\mathbf{x} = \oint_{\partial U_\varepsilon} u \frac{\partial v}{\partial n} - v \frac{\partial u}{\partial n} \, dS$$

Now $\partial U_\varepsilon = \partial U \cup \partial B_\varepsilon$ is the disjoint union of the boundary of U and B_ε , and so $\oint_{\partial U_\varepsilon} = \oint_{\partial U} + \oint_{\partial B_\varepsilon}$. It follows that

$$- \int_{U_\varepsilon} v \Delta u \, d\mathbf{x} - \oint_{\partial U} u \frac{\partial v}{\partial n} - v \frac{\partial u}{\partial n} \, dS = \oint_{\partial B_\varepsilon} u \frac{\partial v}{\partial n} - v \frac{\partial u}{\partial n} \, dS$$

We will be done once we show that the righthand side of the above equation converges to $u(\mathbf{x}_0)$ as $\varepsilon \rightarrow 0$.

Consider, therefore $\oint_{\partial B_\varepsilon} u \frac{\partial v}{\partial n} - v \frac{\partial u}{\partial n} \, dS$. On ∂B_ε the outward unit normal derivative $\frac{\partial}{\partial n}$ is just $-\frac{\partial}{\partial r}$, where $r := \|\mathbf{x} - \mathbf{x}_0\|$, as “outward” from U_ε means “towards the \mathbf{x}_0 ”. It follows that

$$\oint_{\partial B_\varepsilon} u \frac{\partial v}{\partial n} - v \frac{\partial u}{\partial n} \, dS = - \oint_{\partial B_\varepsilon} u \frac{\partial v}{\partial r} + \oint_{\partial B_\varepsilon} v \frac{\partial u}{\partial r} \, dS$$

Let's evaluate the first term of this integral. We consider the case $n \geq 3$, and leave the $n = 2$ -case as an exercise. For $n \geq 3$, we have $\frac{\partial}{\partial r} = -\frac{1}{A_n(1)} \frac{1}{r^{n-1}}$. Also $r = \varepsilon$ on ∂B_ε , so

$$- \oint_{\partial B_\varepsilon} u \frac{\partial v}{\partial r} \, dS = \frac{1}{\varepsilon^{n-1} A_n(1)} \oint_{\partial B_\varepsilon} u \, dS = \frac{1}{A_n(\varepsilon)} \oint_{\partial B_\varepsilon} u \, dS \rightarrow u(\mathbf{x}_0) \quad \text{as } \varepsilon \rightarrow 0$$

For the remaining term, let $M =: \max \left\{ \left| \frac{\partial u(\mathbf{x})}{\partial r} \right| : \|\mathbf{x} - \mathbf{x}_0\| = \varepsilon \right\}$ be the maximum value of $\left| \frac{\partial u}{\partial r} \right|$ on the (compact) surface ∂B_ε . Then we have

$$\oint_{\partial B_\varepsilon} \left| \frac{\partial u}{\partial r} \right| v \, dS = \frac{1}{(n-2)A_n(1)\varepsilon^{n-2}} \oint_{\partial B_\varepsilon} \left| \frac{\partial u}{\partial r} \right| \, dS \leq \frac{1}{(n-2)A_n(1)\varepsilon^{n-2}} M A_n(\varepsilon) = \frac{M}{n-2} \varepsilon$$

and hence

$$\oint_{\partial B_\varepsilon} \frac{\partial u}{\partial r} v(\mathbf{x}) \, dS \rightarrow 0 \quad \text{as } \varepsilon \rightarrow 0$$

We thus see that

$$\oint_{\partial B_\varepsilon} u \frac{\partial v}{\partial n} - v \frac{\partial u}{\partial n} \, dS \rightarrow u(\mathbf{x}_0) + 0 = u(\mathbf{x}_0) \quad \text{as } \varepsilon \rightarrow 0$$

We deduce immediately that

$$- \int_U v \Delta u \, d\mathbf{x} - \oint_{\partial U} u \frac{\partial v}{\partial n} - v \frac{\partial u}{\partial n} \, dS = u(\mathbf{x}_0)$$

and we are done. ◻

Exercise 14.3.6 Finish the proof of the preceding theorem by dealing with the $n = 2$ -case, i.e. show that when u is defined on a bounded open set $U \subseteq \mathbb{R}^2$, then

$$u(\mathbf{x}_0) = \frac{1}{2\pi} \int_U \ln \|\mathbf{x}_0 - \mathbf{x}\| \Delta u(\mathbf{x}) \, d\mathbf{x} + \frac{1}{2\pi} \oint_{\partial U} u(\mathbf{x}) \frac{\partial}{\partial n} \ln \|\mathbf{x}_0 - \mathbf{x}\| - \frac{\partial u}{\partial n} \ln \|\mathbf{x}_0 - \mathbf{x}\| \, ds$$

◻

When u is harmonic, i.e. when $\Delta u = 0$, the above representation simplifies:

Corollary 14.3.7 *Suppose that u is harmonic on a bounded open set $U \subseteq \mathbb{R}^n$. Then*

$$u(\mathbf{x}_0) = \oint_{\partial U} -u(\mathbf{x}) \frac{\partial}{\partial n} \Psi(\mathbf{x} - \mathbf{x}_0) + \Psi(\mathbf{x} - \mathbf{x}_0) \frac{\partial u}{\partial n} dS$$

for all $\mathbf{x}_0 \in U$.

□

It follows then that value of a harmonic function u at a point \mathbf{x}_0 is completely determined by its behaviour “far away”, i.e. by its behaviour on the boundary of any region that contains the point \mathbf{x}_0 . When compared with the mean value property, this result will seem less surprising.

Exercise 14.3.8 Suppose that $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$ is a C^2 -function with compact support (i.e. a twice continuously differentiable function that vanishes outside some compact set K). Show that

$$\phi(0) = - \int \Psi(\mathbf{x}) \Delta \phi(\mathbf{x}) d\mathbf{x}$$

where Ψ is the fundamental solution of Laplace's equation.

[Hint: Choose a bounded open set U which is big enough so that $\phi, \frac{\partial \phi}{\partial n}$ vanish on ∂U and outside U .

□

14.3.3 Green's Functions

Consider now the following Dirichlet problem:

$$\begin{aligned} -\Delta u &= f && \text{in } U \\ u &= g && \text{on } \partial U \end{aligned}$$

where U is a bounded open set with smooth (i.e. C^1) boundary. Fix $\mathbf{x} \in U$, and choose $\varepsilon < 0$ so that the closed ball $B_\varepsilon := \bar{B}(\mathbf{x}, \varepsilon) \subseteq U$, and let $U_\varepsilon = U - B_\varepsilon$.

In order to use the representation in Theorem 14.3.5, i.e.

$$u(\mathbf{x}) = - \int_U \Psi(\mathbf{y} - \mathbf{x}) \Delta u(\mathbf{y}) d\mathbf{y} + \oint_{\partial U} \Psi(\mathbf{y} - \mathbf{x}) \frac{\partial u}{\partial n} - u(\mathbf{y}) \frac{\partial}{\partial n} \Psi(\mathbf{y} - \mathbf{x}) dS(\mathbf{y}) \quad (*)$$

we need to know Δu in U — which we do — and also the normal derivative $\frac{\partial u}{\partial n}$ on ∂U — which we don't. To get around this we seek, for this \mathbf{x} , a *corrector function* $h^\mathbf{x} = h^\mathbf{x}(\mathbf{y})$ which solves an auxiliary boundary value problem:

$$\begin{aligned} \Delta h^\mathbf{x} &= 0 && \text{in } U \\ h^\mathbf{x}(\mathbf{y}) &= \Psi(\mathbf{y} - \mathbf{x}) && \text{on } \partial U \end{aligned}$$

Assuming that such a function $h^\mathbf{x}$ can be found, we see that, by Green's formula,

$$- \int_U h^\mathbf{x}(\mathbf{y}) \Delta u(\mathbf{y}) d\mathbf{y} = \oint_{\partial U} u(\mathbf{y}) \frac{\partial h^\mathbf{x}}{\partial n}(\mathbf{y}) - h^\mathbf{x}(\mathbf{y}) \frac{\partial u}{\partial n}(\mathbf{y}) dS(\mathbf{y})$$

which implies that

$$0 = \oint_{\partial U} u(\mathbf{y}) \frac{\partial h^{\mathbf{x}}}{\partial n}(\mathbf{y}) - \Psi(\mathbf{y} - \mathbf{x}) \frac{\partial u}{\partial n}(\mathbf{y}) dS(\mathbf{y}) \quad (\dagger)$$

Adding (*), (\dagger), we obtain

$$u(\mathbf{x}) = - \int_U (\Psi(\mathbf{y} - \mathbf{x}) - h^{\mathbf{x}}(\mathbf{y})) \Delta u(\mathbf{y}) d\mathbf{y} - \oint_{\partial U} u(\mathbf{y}) \frac{\partial}{\partial n} (\Psi(\mathbf{y} - \mathbf{x}) - h^{\mathbf{x}}(\mathbf{y})) dS(\mathbf{y})$$

We therefore define the *Green's function* $G(\mathbf{x}, \mathbf{y})$ for the region U by

$$G(\mathbf{x}, \mathbf{y}) := \Psi(\mathbf{y} - \mathbf{x}) - h^{\mathbf{x}}(\mathbf{y})$$

and obtain

$$u(\mathbf{x}) = - \int_U G(\mathbf{x}, \mathbf{y}) \Delta u(\mathbf{y}) d\mathbf{y} - \oint_{\partial U} u(\mathbf{y}) \frac{\partial}{\partial n} G(\mathbf{x}, \mathbf{y}) dS(\mathbf{y})$$

Noting that we require $\Delta u = f$ in U and $u = g$ on ∂U , we see that we have proved the following theorem:

Theorem 14.3.9 (Green's function solution of Laplace's equation)

Suppose that $u \in C^2(\bar{U})$ is a solution of the inhomogeneous Dirichlet problem

$$\begin{aligned} -\Delta u &= f && \text{in } U \\ u &= g && \text{on } \partial U \end{aligned}$$

where U is a bounded open set with C^1 boundary. The Green's function for U is defined to be

$$G(\mathbf{x}, \mathbf{y}) = \Psi(\mathbf{y} - \mathbf{x}) - h^{\mathbf{x}}(\mathbf{y})$$

where, for $\mathbf{x} \in U$, the function $h^{\mathbf{x}}(\mathbf{y})$ is a solution to the BVP

$$\begin{aligned} \Delta h^{\mathbf{x}} &= 0 && \text{in } U \\ h^{\mathbf{x}}(\mathbf{y}) &= \Psi(\mathbf{y} - \mathbf{x}) && \text{on } \partial U \end{aligned}$$

The solution $u(\mathbf{x})$ then has representation

$$u(\mathbf{x}) = \int_U f(\mathbf{y}) G(\mathbf{x}, \mathbf{y}) d\mathbf{y} - \oint_{\partial U} g(\mathbf{y}) \frac{\partial}{\partial n} G(\mathbf{x}, \mathbf{y}) dS(\mathbf{y})$$

(where the normal derivative $\frac{\partial G}{\partial n}$ is understood to be with respect to the \mathbf{y} -variable.)

□

We have therefore — in principle — solved the inhomogeneous Laplace equation: We just need to compute the Green's function for the appropriate region. However, this is usually a non-trivial task. We will give some examples shortly. For the moment, we want to get a better understanding of the meaning of the Green's function.

14.3.4 Properties of Green's functions

Proposition 14.3.10 *The Green's function $G(\mathbf{x}_0, \mathbf{x})$ for a bounded open set U is the unique $C^2(\bar{U})$ -function such that, fixing \mathbf{x}_0 , and regarding $G^{\mathbf{x}_0}(\mathbf{x}) := G(\mathbf{x}_0, \mathbf{x})$ as a function of \mathbf{x} :*

- (i) $\Delta G = 0$ in U (i.e. $G(\mathbf{x}_0, \mathbf{x})$ is harmonic in U), except at the point $\mathbf{x} = \mathbf{x}_0$.
- (ii) $G(\mathbf{x}_0, \mathbf{x}) - \Psi(\mathbf{x} - \mathbf{x}_0)$ is finite at \mathbf{x}_0 , and is harmonic.
- (iii) $G(\mathbf{x}_0, \mathbf{x}) = 0$ for $\mathbf{x} \in \partial U$.

□

Exercise 14.3.11 We prove Propn. 14.3.10

- (a) First prove uniqueness, i.e. that if G_1, G_2 are two functions satisfying (i)–(iii), then $G_1 = G_2$.
- (b) It therefore remains to show that the Green's function for a region satisfies (i)–(iii). Do so.
[Recall that $G(\mathbf{x}_0, \mathbf{x}) := \Psi(\mathbf{x} - \mathbf{x}_0) - h^{\mathbf{x}_0}(\mathbf{x})$ where the corrector function $h^{\mathbf{x}_0}$ is the solution of the auxiliary BVP

$$\begin{aligned} \Delta h^{\mathbf{x}_0} &= 0 && \text{in } U \\ h^{\mathbf{x}_0}(\mathbf{x}) &= \Psi(\mathbf{x} - \mathbf{x}_0) && \text{on } \partial U \end{aligned}$$

(You may assume that this BVP has a solution; its existence follows from existence theory for the Dirichlet problem.)]

□

Remarks 14.3.12 It is clear from the preceding exercise that the corrector function $h(\mathbf{x}_0, \mathbf{x}) := h^{\mathbf{x}_0}(\mathbf{x})$ is the unique harmonic function which subtracts (or adds) just the right amount to the fundamental solution $\Psi(\mathbf{x}_0, \mathbf{x}) := \Psi(\mathbf{x} - \mathbf{x}_0)$, for each \mathbf{x}_0 , to ensure that $G := \Psi - h$ satisfies the correct boundary conditions.

□

Proposition 14.3.13 *The Green's function $G(\mathbf{x}, \mathbf{y})$ for a bounded open set U is symmetric in \mathbf{x} and \mathbf{y} , i.e.*

$$G(\mathbf{x}, \mathbf{y}) = G(\mathbf{y}, \mathbf{x})$$

Proof: Fix $\mathbf{x}_0, \mathbf{y}_0 \in U$. We will show that $G(\mathbf{x}_0, \mathbf{y}_0) = G(\mathbf{y}_0, \mathbf{x}_0)$. If $\mathbf{x}_0 = \mathbf{y}_0$, there is nothing to prove, so assume that $\mathbf{x}_0 \neq \mathbf{y}_0$, and define

$$u(\mathbf{x}) := G(\mathbf{x}_0, \mathbf{x}) \quad v(\mathbf{x}) := G(\mathbf{y}_0, \mathbf{x})$$

u is singular at \mathbf{x}_0 and v at \mathbf{y}_0 , so we isolate the singularities in small closed balls $A_\varepsilon := \bar{B}(\mathbf{x}_0, \varepsilon)$, $B_\varepsilon := \bar{B}(\mathbf{y}_0, \varepsilon)$, where $\varepsilon > 0$ is chosen sufficiently small to ensure that the balls are

disjoint and contained in U . Now define $U_\varepsilon := U - (A_\varepsilon \cup B_\varepsilon)$. Note that u, v are harmonic in U_ε and that $u(\mathbf{x}) = 0 = v(\mathbf{x})$ for $\mathbf{x} \in \partial U$. By Green's Theorem

$$\int_{U_\varepsilon} u \Delta v - v \Delta u \, d\mathbf{x} = \oint_{\partial U} u \frac{\partial v}{\partial n} - v \frac{\partial u}{\partial n} \, dS + I_\varepsilon + J_\varepsilon$$

where

$$I_\varepsilon = \oint_{\partial A_\varepsilon} u \frac{\partial v}{\partial n} - v \frac{\partial u}{\partial n} \, dS \quad J_\varepsilon = - \oint_{\partial B_\varepsilon} v \frac{\partial u}{\partial n} - u \frac{\partial v}{\partial n} \, dS$$

Since u, v are harmonic in U_ε and zero on ∂U , we see that $I_\varepsilon + J_\varepsilon = 0$ for all $\varepsilon > 0$, so that also

$$\lim_{\varepsilon \rightarrow 0} I_\varepsilon + \lim_{\varepsilon \rightarrow 0} J_\varepsilon = 0$$

We now compute $\lim_{\varepsilon \rightarrow 0} I_\varepsilon$: Since $v(\mathbf{x})$ is smooth (at least C^2) near \mathbf{x}_0 , its derivatives are continuous, hence bounded near \mathbf{x}_0 . As $u(\mathbf{x}) := G(\mathbf{x}, \mathbf{x}_0) := \Psi(\mathbf{x} - \mathbf{x}_0) - h^{\mathbf{x}_0}(\mathbf{x})$, and as $h^{\mathbf{x}_0}$ is smooth, hence bounded near \mathbf{x}_0 , it follows that

$$\left| \oint_{\partial A_\varepsilon} u \frac{\partial v}{\partial n} \, dS \right| \leq \left| \Psi(\varepsilon) \oint_{\partial A_\varepsilon} \frac{\partial v}{\partial n} \, dS - \oint_{\partial A_\varepsilon} h^{\mathbf{x}_0} \frac{\partial v}{\partial n} \, dS \right| \rightarrow 0 \quad \text{as } \varepsilon \rightarrow 0$$

The other term of I_ε is

$$- \oint_{\partial A_\varepsilon} v \frac{\partial u}{\partial n} \, dS \rightarrow v(\mathbf{x}_0) \quad \text{as } \varepsilon \rightarrow 0$$

Hence $\lim_{\varepsilon \rightarrow 0} I_\varepsilon = v(\mathbf{x}_0)$. In the same way, $\lim_{\varepsilon \rightarrow 0} J_\varepsilon = -u(\mathbf{y}_0)$. As $\lim_{\varepsilon \rightarrow 0} I_\varepsilon + \lim_{\varepsilon \rightarrow 0} J_\varepsilon = 0$ we see that $v(\mathbf{x}_0) = u(\mathbf{y}_0)$ and so

$$\mathbf{G}(\mathbf{x}_0, \mathbf{y}_0) = v(\mathbf{x}_0) = u(\mathbf{y}_0) = G(\mathbf{y}_0, \mathbf{x}_0)$$

Since $\mathbf{x}_0, \mathbf{y}_0$ were arbitrary, the result follows. ◻

One way to interpret the Green's function is that it is the kernel of a linear operator which inverts the PDE. Consider again Theorem 14.3.9, and note that the Green's function representation of the solution to the *Poisson* problem with homogeneous Dirichlet boundary conditions

$$\begin{aligned} -\Delta v &= f & \text{in } U \\ v &= 0 & \text{on } \partial U \end{aligned}$$

is particularly simple:

$$v(\mathbf{x}) = \int_U G(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) \, d\mathbf{y}$$

Define a linear operator (on functions) L_P by

$$L_P[f](\mathbf{x}) = \int_U G(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) \, d\mathbf{y}$$

L_P is said to be an *integral operator* with kernel G . It is immediately clear that — just like integration is an inverse to differentiation — the integral operator L_P is a kind of inverse to the differential operator $-\Delta$:

$$-\Delta v = f \quad \Longleftrightarrow \quad v = L_P[f]$$

In the same way the *Dirichlet* problem with non-homogeneous boundary conditions

$$\begin{aligned} -\Delta w &= 0 & \text{in } U \\ w &= g & \text{on } \partial U \end{aligned}$$

has representation

$$w(\mathbf{x}) = - \oint_{\partial U} g(\mathbf{y}) \frac{\partial}{\partial n} G(\mathbf{x}, \mathbf{y}) dS(\mathbf{y})$$

The integral operator L_D with kernel $-\frac{\partial}{\partial n} G(\mathbf{x}, \mathbf{y})$ inverts the Dirichlet problem.

Now note that if v, w are solutions to

$$\begin{aligned} -\Delta v &= f & -\Delta w &= 0 & \text{in } U \\ v &= 0 & w &= g & \text{on } \partial U \end{aligned}$$

then, by linearity, it follows that $u = v + w$ is a solution to a general Dirichlet–Poisson problem

$$\begin{aligned} -\Delta u &= f & \text{in } U \\ u &= g & \text{on } \partial U \end{aligned}$$

As the solution to such a problem is unique, it follows that u is *the* solution, i.e. the Dirichlet–Poisson problem can be regarded as two separate problems, a Poisson problem and a Dirichlet problem. Rewrite the problem suggestively as

$$\begin{pmatrix} -\Delta \\ I \end{pmatrix} u = \begin{pmatrix} f \\ g \end{pmatrix} \quad \text{on } \begin{pmatrix} U \\ \partial U \end{pmatrix}$$

and define L by

$$L \begin{pmatrix} f \\ g \end{pmatrix} := L_P[f] + L_D[g]$$

and we see that

$$\begin{pmatrix} -\Delta \\ I \end{pmatrix} u = \begin{pmatrix} f \\ g \end{pmatrix} \quad \Longleftrightarrow \quad u = L \begin{pmatrix} f \\ g \end{pmatrix}$$

i.e. L applied to the data gives the solution of the BVP.

Another way to see the Green's function is to note that

$$\int_U f(\mathbf{y}) G(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) d\mathbf{y}$$

is a kind of weighted sum of terms $f(\mathbf{y})$, weighted by $G(\mathbf{x}, \mathbf{y})$. Thus the solution $v(\mathbf{x})$ of the Poisson problem may be regarded as a kind of “average” of the data f , weighted by the Green's function. The Green's function $G(\mathbf{x}, \mathbf{y})$ determines the “strength” of the contribution of $f(\mathbf{y})$ to the solution at the point \mathbf{x} . The fact that $G(\mathbf{x}, \mathbf{y}) = G(\mathbf{y}, \mathbf{x})$ is like Newton's First Law: The strength of the effect of \mathbf{y} at \mathbf{x} is equal to the strength of the effect of \mathbf{x} at \mathbf{y} . Analogous remarks can be made for the Dirichlet problem: The solution $w(\mathbf{x})$ is a kind of weighted average of the data g , weighted by the *inward* normal derivative $-\frac{\partial}{\partial n} G(\mathbf{x}, \mathbf{y})$ of the Green's function on the boundary.

We will further develop this interpretation when we discuss Green's functions for the heat equation.

Exercise 14.3.14 We will solve the one-dimensional Dirichlet–Poisson problem

$$\begin{aligned} -u''(x) &= f(x) & 0 < x < l \\ u(0) &= u(l) = 0 \end{aligned}$$

(a) Show that the fundamental solution of the one-dimensional Laplace equation is

$$\Psi(x) = \frac{1}{2}|x|$$

(b) Hence solve the auxiliary problem for the corrector function $\frac{d^2}{dy^2}h^x(y) = 0, h^x(0) = \Psi(-x), h^x(l) = \Psi(l-x)$ to show that Green's function $G(x, y)$ for the open set $U = (0, l)$ is given by

$$G(x, y) = \begin{cases} \frac{x(y-l)}{l} & \text{if } y \geq x \\ \frac{y(x-l)}{l} & \text{if } y \leq x \end{cases}$$

[Hint: Recall that h^x must be smooth in the whole region U to deduce that $h^x(y) = a(x)y + b(x)$ for some functions a, b .]

(c) Note that G is symmetric (which is always the case), and that G is *not* singular (which is hardly ever the case).

(d) Because G is not singular, differentiation under the integral sign is permitted. Verify that

$$v(x) = \int_0^l G(x, y)f(y) dy$$

solves the given BVP.

[Hint: Recall (or if you don't recall, prove) that

$$\frac{d}{dx} \int_{a(x)}^{b(x)} g(x, y) dy = g(x, b(x)) - g(x, a(x)) + \int_{a(x)}^{b(x)} \frac{\partial}{\partial x} g(x, y) dy$$

and split the integral: $\int_0^l = \int_0^x + \int_x^l$]

□

14.4 Green's Functions: Examples and Exercises

14.4.1 Green's function for the half-space

Consider the *half-space*

$$\begin{aligned} \mathbb{H}^n &:= \{(x_1, \dots, x_n) \in \mathbb{R}^n : x_n > 0\} \\ \partial\mathbb{H}^n &= \{(x_1, \dots, x_n) \in \mathbb{R}^n : x_n = 0\} \end{aligned}$$

We want to determine the Green's function for this region. Note, however, that the half-space is open, but unbounded, so that the results of the previous section may not hold. However, if a “boundary condition at $+\infty$ ” holds, which asserts that the functions and their derivatives tend to 0 as $\|\mathbf{x}\| \rightarrow \infty$, then the results can be proven valid. In practice, one

proceeds as follows: Assume that the results of the previous section do hold, in order to find a candidate G . Then show that this G has the required properties.

Recall that, by Thm. 14.3.10 G is the *unique* $C^2(\bar{\mathbb{H}}^n)$ -function such that, fixing \mathbf{x}_0 , and regarding $G^{\mathbf{x}_0}(\mathbf{x}) := G(\mathbf{x}_0, \mathbf{x})$ as a function of \mathbf{x} :

- (i) G is harmonic in \mathbb{H}^n , except possibly at the point $\mathbf{x} = \mathbf{x}_0$.
- (ii) The difference $G(\mathbf{x}_0, \mathbf{x}) - \Psi(\mathbf{x} - \mathbf{x}_0)$ is *finite* at \mathbf{x}_0 , and is harmonic.
- (iii) $G(\mathbf{x}_0, \mathbf{x}) = 0$ for $\mathbf{x} \in \partial\mathbb{H}^n$.

Note that the function $\Psi^{\mathbf{x}_0}(\mathbf{x}) := \Psi(\mathbf{x} - \mathbf{x}_0)$ satisfies (i) and (ii), but not the boundary condition (iii). We now use a trick, called the *method of images*, to ensure that the boundary condition $G = 0$ is satisfied by exploiting symmetry. (This method will be employed again when we solve the Black–Scholes PDE to price barrier options.) Given a point $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$, define its reflection in the plane $\partial\mathbb{H}^n$ by

$$\mathbf{x}^* := (x_1, \dots, x_{n-1}, -x_n)$$

and note that $\mathbf{x} = \mathbf{x}^*$ if and only if $\mathbf{x} \in \partial\mathbb{H}^n$. Suppose that u is harmonic in \mathbb{R}^n so that $u(\mathbf{x}) = -u(\mathbf{x}^*)$ for all \mathbf{x} . By continuity, if $\mathbf{x} \in \partial\mathbb{H}^n$, we see that also that $u(\mathbf{x}) = u(\mathbf{x}^*)$, and hence that $u(\mathbf{x}) = 0$ for all $\mathbf{x} \in \partial\mathbb{H}^n$. Thus: In order to find a harmonic function which satisfies the boundary condition $u = 0$ on $\partial\mathbb{H}^n$, it suffices to find a harmonic function satisfying the symmetry relation

$$u(\mathbf{x}) = -u(\mathbf{x}^*)$$

This immediately suggests a candidate for the Green's function, namely

$$G(\mathbf{x}_0, \mathbf{x}) := \Psi(\mathbf{x} - \mathbf{x}_0) - \Psi(\mathbf{x} - \mathbf{x}_0^*)$$

i.e. the corrector function $h^{\mathbf{x}_0}$ reflects the singularity from \mathbf{x}_0 to \mathbf{x}_0 so that the contributions add up to zero on the boundary. We now check that this G satisfies (i)–(iii):

- (i) G is the difference of harmonic functions, and therefore harmonic, as the Laplace operator is linear.
- (ii) $G(\mathbf{x}_0, \mathbf{x}) - \Psi(\mathbf{x} - \mathbf{x}_0) = -\Psi(\mathbf{x} - \mathbf{x}_0^*)$ is finite at \mathbf{x}_0 — the singularity is at $\mathbf{x}_0^* \notin \mathbb{H}^n$.
- (iii) Recall that the fundamental Ψ is radial, i.e. that $\Psi(\mathbf{x}) = \Psi(r)$ depends only on $r = \|\mathbf{x}\|$. Since

$$\|\mathbf{x} - \mathbf{x}_0\| = \|\mathbf{x} - \mathbf{x}_0^*\| \quad \text{when } \mathbf{x} \in \partial\mathbb{H}^n$$

it follows that $G(\mathbf{x}_0, \mathbf{x}) = \Psi(r) - \Psi(r) = 0$, where $r := \|\mathbf{x} - \mathbf{x}_0\| = \|\mathbf{x} - \mathbf{x}_0^*\|$.

In order to use the representation theorem 14.3.9, we need to know the normal derivative $\frac{\partial G^{\mathbf{x}_0}}{\partial n}$. The unit outward normal to $\partial\mathbb{H}^n$ is clearly a constant vector field $\mathbf{n} = (0, \dots, 0, -1)$, so that $\frac{\partial}{\partial n} = -\frac{\partial}{\partial x_n}$. It is straightforward to compute that

$$\frac{\partial G}{\partial x_n} = -\frac{1}{A_n(1)} \left[\frac{x_n - x_{0n}}{\|\mathbf{x} - \mathbf{x}_0\|^n} - \frac{x_n - x_{0n}^*}{\|\mathbf{x} - \mathbf{x}_0^*\|^n} \right]$$

When $\mathbf{x} \in \partial\mathbb{H}^n$, we have $\|\mathbf{x} - \mathbf{x}_0\| = \|\mathbf{x} - \mathbf{x}_0^*\|$ and $x_{0n} = 0 = x_{0n}^*$, and so

$$\frac{\partial G}{\partial n}(\mathbf{x}_0, \mathbf{x}) = -\frac{2x_n}{A_n(1)\|\mathbf{x} - \mathbf{x}_0\|^n}$$

Exercise 14.4.1 Perform the computations for $\frac{\partial G}{\partial x_n}$ above, for both the cases $n = 2$ and $n \geq 3$. □

Thus, from Theorem 14.3.9 we expect that

Theorem 14.4.2 (Poisson Formula) *Define the Poisson kernel for $\partial\mathbb{H}^n$ by*

$$K(\mathbf{x}, \mathbf{y}) := \frac{2x_n}{A_n(1) \|\mathbf{x} - \mathbf{y}\|^n}$$

Then the Dirichlet–Poisson problem

$$\begin{aligned} \Delta u &= f && \text{in } \mathbb{H}^n \\ u &= g && \text{on } \partial\mathbb{H}^n \end{aligned}$$

has solution

$$u(\mathbf{x}) = \int_{\mathbb{H}^n} G(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) \, d\mathbf{y} + \int_{\partial\mathbb{H}^n} K(\mathbf{x}, \mathbf{y}) g(\mathbf{y}) \, d\mathbf{y}$$

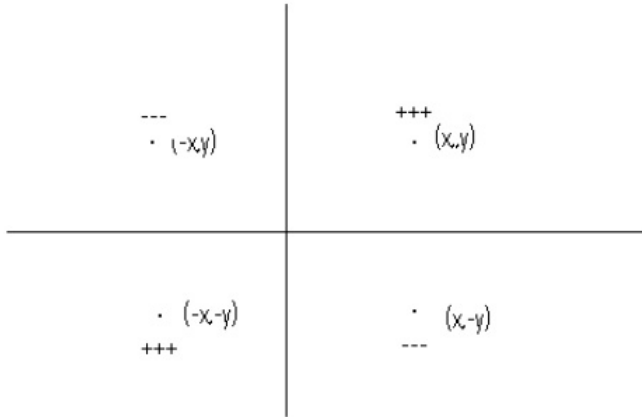
□

Exercise 14.4.3 Use the method of images to determine the Green's function for the positive quadrant

$$U := \{(x, y) \in \mathbb{R}^2 : x > 0, y > 0\}$$

[Hint: Fix a point $(\xi, \eta) \in \mathbb{R}^2$. We seek a function $G(\xi, \eta; x, y)$ of (x, y) which is harmonic (in x, y) and has the properties that $G(\xi, \eta; x, y) - \Psi(x - \xi, y - \eta)$ is harmonic and finite in U , and such that

$$G(\xi, \eta; 0, x) = 0 = G(\xi, \eta; x, 0) \quad \text{for all } x \geq 0, y \geq 0$$



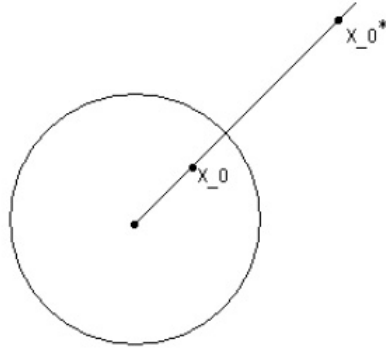
Show that

$$\begin{aligned} G(\xi, \eta; x, y) &:= \Psi(x - \xi, y - \eta) - \Psi(x + \xi, y - \eta) - \Psi(x - \xi, y + \eta) + \Psi(x + \xi, y + \eta) \\ &= \frac{1}{4\pi} \ln \left[\frac{(\xi - x)^2 + (\eta - y)^2}{(\xi + x)^2 + (\eta - y)^2} \cdot \frac{(\xi + x)^2 + (\eta + y)^2}{(\xi - x)^2 + (\eta + y)^2} \right] \end{aligned}$$

does the trick. □

14.4.2 Green's function for the ball

Do the following exercise:



Exercise 14.4.4 Let $U : B(0, R) \subseteq \mathbb{R}^n$ be the open ball of radius R centered at the origin. We seek the Green's function G for the bounded region U .

(a) Again, we use a kind of reflection (an inversion, really) in the boundary. Given a point \mathbf{x}_0 , we want the “reflection” \mathbf{x}_0^* to satisfy two properties:

- (i) $\mathbf{x}_0, \mathbf{x}_0^*$ lie on a line through the origin; and,
- (ii) $\|\mathbf{x}_0\| \cdot \|\mathbf{x}_0^*\| = R^2$, which ensures that the “reflection” of a point on the boundary is itself.

Show that if $\mathbf{x}_0 \neq 0$, the “reflection” of \mathbf{x}_0 in ∂U is given by

$$\mathbf{x}_0^* := \frac{R^2 \mathbf{x}_0}{\|\mathbf{x}_0\|^2}$$

(b) Now for a little geometry: For an arbitrary $\mathbf{x} \in \mathbb{R}^n$, define

$$r := \|\mathbf{x} - \mathbf{x}_0\| \quad r^* := \|\mathbf{x} - \mathbf{x}_0^*\| \quad \tilde{r} := \|\mathbf{x}^* - \mathbf{x}_0^*\|$$

Show that

$$\tilde{r} = \frac{R^2}{\|\mathbf{x}_0\| \|\mathbf{x}\|} r$$

$$[\text{Hint: } \tilde{r}^2 = (\mathbf{x}^* - \mathbf{x}_0^*) \cdot (\mathbf{x}^* - \mathbf{x}_0^*) = \cdots = \frac{R^4}{\|\mathbf{x}\|^2 \|\mathbf{x}_0\|^2} r^2.]$$

(c) First, consider the case where $\mathbf{x} \neq 0$. As before the fundamental solution Ψ is a candidate for the Green's function G , but it fails to satisfy the boundary condition, namely that $G(\mathbf{x}_0, \mathbf{x}) = 0$ for $\mathbf{x} \in \partial B(0, R)$. To ensure that $G = 0$ is satisfied, we need to invert the singularity. The contribution of the point \mathbf{x}_0^* at a point \mathbf{x} on the boundary must be just enough to cancel the effect of $\Psi(\mathbf{x} - \mathbf{x}_0) = \Psi(r)$. We thus attempt a solution of the form

$$G^{\mathbf{x}_0}(\mathbf{x}) := \Psi(r) - k\Psi(r^*)$$

where the constant k (which may depend on \mathbf{x}_0 , but not on \mathbf{x}) is chosen so that $G^{\mathbf{x}_0}(\mathbf{x}) = 0$ when $\mathbf{x} \in \partial B(0, R)$. Show that we require $k = \left(\frac{r^*}{R}\right)^{n-2}$. By observing that $r^* = \tilde{r}$ when $\mathbf{x} \in \partial B(0, R)$, conclude that

$$k = \left(\frac{R}{\|\mathbf{x}_0\|}\right)^{n-2}$$

which is indeed independent of \mathbf{x} .

(d) Show that

$$k\Psi(r^*) = \Psi\left(\frac{\|\mathbf{x}_0\|r^*}{R}\right) \quad \text{i.e. that} \quad k\Psi(\mathbf{x} - \mathbf{x}_0^*) = \Psi\left(\frac{\|\mathbf{x}_0\|}{R}(\mathbf{x} - \mathbf{x}_0^*)\right)$$

Conclude that the candidate Green's function is given by

$$G(\mathbf{x}_0, \mathbf{x}) := \Psi(\mathbf{x} - \mathbf{x}_0) - \Psi\left(\frac{\|\mathbf{x}_0\|}{R}(\mathbf{x} - \mathbf{x}_0^*)\right) \quad \text{for } \mathbf{x}_0 \neq 0$$

(e) Now show that G satisfies the requirements of a Green's function.

(f) We still need to find the Green's function $G(\mathbf{x}_0, \mathbf{x})$ in case $\mathbf{x} = 0$. Show that

$$G(0, \mathbf{x}) = \Psi(\mathbf{x}) - \Psi(R)$$

□

Exercise 14.4.5 We solve the Dirichlet problem

$$\begin{aligned} \Delta u &= 0 & \text{in } B(0, R) \\ u &= g & \text{on } \partial B(0, R) \end{aligned}$$

Assume first that $R = 1$. By the representation theorem 14.3.9, we need merely find the outward normal derivative $\frac{\partial G}{\partial n}$ on $\partial B(0, R)$. Show that

$$\frac{\partial}{\partial n} = \frac{\partial}{\partial r} = \sum_{i=1}^n x_i \frac{\partial}{\partial x_i}$$

Then show that

$$\frac{\partial \Psi}{\partial x_i}(\mathbf{x} - \mathbf{x}_0) = -\frac{1}{A_n(1)} \frac{x_i - x_{0i}}{\|\mathbf{x} - \mathbf{x}_0\|^n}$$

and that

$$\begin{aligned} \frac{\partial \Psi}{\partial x_i}(\|\mathbf{x}_0\|(\mathbf{x} - \mathbf{x}_0^*)) &= -\frac{1}{A_n(1)} \frac{x_i - x_{0i}^*}{\|\mathbf{x}_0\|^{n-2} \|\mathbf{x} - \mathbf{x}_0^*\|^n} \\ &= -\frac{1}{A_n(1)} \frac{x_i \|\mathbf{x}_0\|^2 - x_{0i}}{\|\mathbf{x} - \mathbf{x}_0\|^n} \end{aligned}$$

(Use the relation $\tilde{r} = Rr/\|\mathbf{x}_0\|$ from the previous problem. Put $R = 1$ and note that $r^* = \tilde{r}$ when $\mathbf{x} \in \partial B(0, 1)$. Furthermore, by definition $r^* := \|\mathbf{x} - \mathbf{x}_0\|$.) Sum to show that

$$\frac{\partial G}{\partial n}(\mathbf{x}_0, \mathbf{x}) = -\frac{1}{A_n(1)} \frac{1 - \|\mathbf{x}_0\|^2}{\|\mathbf{x} - \mathbf{x}_0\|^n}$$

Thm. 14.3.9 now implies that

$$u(\mathbf{x}) = \frac{1 - \|\mathbf{x}\|^2}{A_n(1)} \oint_{\partial B(0,1)} \frac{g(\mathbf{y})}{\|\mathbf{y} - \mathbf{x}\|^n} dS(\mathbf{y})$$

If $R \neq 1$, use the change of variable $\tilde{u}(\mathbf{x}) := u(R\mathbf{x})$, $\tilde{g}(\mathbf{x}) := g(R\mathbf{x})$ to show that

$$u(\mathbf{x}) = \frac{R^2 - \|\mathbf{x}\|^2}{A_n(1)R} \oint_{\partial B(0,R)} \frac{g(\mathbf{y})}{\|\mathbf{y} - \mathbf{x}\|^n} dS(\mathbf{y}) \quad \text{for } \mathbf{x} \in B(0, R)$$

i.e.

$$u(\mathbf{x}) = \oint_{\partial B(0,R)} K(\mathbf{x}, \mathbf{y}) g(\mathbf{y}) dS(\mathbf{y})$$

where $K(\mathbf{x}, \mathbf{y}) := \frac{R^2 - \|\mathbf{x}\|^2}{A_n(1)R} \frac{1}{\|\mathbf{x} - \mathbf{y}\|^n}$ is *Poisson's kernel* for the ball $B(0, R)$.

□

Exercise 14.4.6 Start with the solution of the Dirichlet problem for the ball of radius R , given in Exercise 14.4.5, and restrict to the two-dimensional case. Convert to polar coordinates in \mathbb{R}^2 to deduce that the solution of the Dirichlet problem with boundary data g is

$$u(r_0, \theta_0) = \int_0^{2\pi} K(r, \theta; R, \phi) g(\phi) d\phi$$

where the Poisson kernel for a ball of radius R is given by

$$K(r, \theta; R, \phi) = \frac{R^2 - r^2}{R^2 - 2Rr \cos(\theta - \phi) + r^2}$$

This is *Poisson's formula* (in \mathbb{R}^2).

□

Exercise 14.4.7 Use Poisson's formula, obtained in Exercise 14.4.6 to show that the solution of

$$\begin{aligned} \Delta u &= 0 && \text{in } B(0, \sqrt{6}) \\ u &= y + y^2 && \text{on } \partial B(0, \sqrt{6}) \end{aligned}$$

is given by

$$u(x, y) = 3 + y + \frac{1}{2}(y^2 - x^2)$$

[Hint: Convert to polar coordinates and show first that $u(r, \theta) = 3 - r^2 \cos^2 \theta + \frac{r^2}{2} + r \sin \theta$.]

□

Chapter 15

The Heat Equation

We study the heat equation

$$u_t - \frac{1}{2}\Delta u = 0$$

and its non-homogeneous counterpart

$$u_t - \frac{1}{2}\Delta u = f$$

subject to appropriate initial and boundary conditions. Here $u = u(t, \mathbf{x})$, and where $t > 0$ is time (which plays a special role) and $\mathbf{x} = (x_1, \dots, x_n)$ is a spatial variable, $\mathbf{x} \in U$ for some open set $U \subseteq \mathbb{R}^n$. The Laplacian operator Δ refers to the spatial part only, i.e. $\Delta = \sum_{j=1}^n \frac{\partial^2}{\partial x_j^2}$.

Exercise 15.0.8 Suppose that $u(t, \mathbf{x})$ is a solution of the heat equation, that $\mathbf{x}_0 \in \mathbb{R}^n$ and that O is an orthogonal $n \times n$ -matrix. Show that the functions

$$v(t, \mathbf{x}) := u(t, \mathbf{x} - \mathbf{x}_0) \quad w(t, \mathbf{x}) := u(t, O\mathbf{x})$$

are also solutions of the heat equation. Conclude that the heat equation is invariant under spatial translations and rotations.

□

15.1 Separation of Variables

Let $L[\]$ denote a linear partial differential operator, i.e. one so that $L[f] = 0$ is a linear PDE. It is then easy to see that the set of solutions

$$\{f : L[f] = 0\}$$

is a linear space. For if f, g are solutions and α, β are scalars, then

$$L[\alpha f + \beta g] = \alpha L[f] + \beta L[g] = 0$$

and one might hazard that if $f = \sum_{n=1}^{\infty} c_n u_n$, and $L[u_n] = 0$ for all n , then also $L[f] = 0$. The *separation of variables* technique works by writing the solution $f(\mathbf{x})$ as a product $f(x^1, \dots, x^n) = X^1(x^1) \cdots X^n(x^n)$ of n functions X^1, \dots, X^n of just one variable each, and then adding these up to ensure that the correct initial- and/or boundary conditions are satisfied.

Example 15.1.1 Suppose that we want to find the solution $u(t, x)$ which satisfies the following conditions:

$$\begin{array}{lll} \text{PDE} & u_t = \alpha^2 u_{xx} & 0 < x < L, \quad 0 < t < \infty \\ \text{BC} & u(t, 0) = 0 = u(t, L) & 0 < t < \infty \\ \text{IC} & u(0, x) = \phi(x) & \end{array}$$

This corresponds to the following physical situation: $u(t, x)$ is the temperature at time t at point x , in a laterally insulated rod of length L , both of whose sides are kept at a temperature of 0.

To solve this problem, we attempt to *separate* the time- and spatial variables, and hypothesize that the solution has the form

$$u(t, x) = T(t)X(x)$$

Substituting this into the PDE yields

$$T'(t)X(x) = \alpha^2 T(t)X''(x)$$

which is equivalent to

$$\frac{T'(t)}{\alpha^2 T(t)} = \frac{X''(x)}{X(x)} \quad (*)$$

Now since the lefthand side of $(*)$ is independent of x , and since the righthand side equals the lefthand side, the righthand side is independent of x also! Similarly, both sides are independent of t (because the righthand side is independent of t). Thus both sides are constant, i.e. there is a constant k such that

$$\frac{T'(t)}{\alpha^2 T(t)} = k = \frac{X''(x)}{X(x)}$$

The PDE has now been reduced to two ODE's:

$$\begin{aligned} T' - k\alpha^2 T &= 0 \\ X'' - kX &= 0 \end{aligned}$$

Solving first for T , we see that

$$T(t) = Ce^{k\alpha^2 t} \quad C \text{ constant}$$

This blows up if $k > 0$, which is “unphysical”, and we therefore assume that $k = -\lambda^2 \leq 0$. Solving the ODE for X , we get

$$X(x) = A \cos \lambda x + B \sin \lambda x \quad A, B \text{ constant}$$

so that

$$u(t, x) = e^{-\lambda^2 \alpha^2 t} [A \cos \lambda x + B \sin \lambda x]$$

This gives us a solution to the problem for every value of λ, A, B .

We now impose the boundary conditions: The requirement that $u(t, 0) = 0$ for all t is easily seen to imply that $A = 0$. The requirement that $u(t, L) = 0$ for all t then implies that

$\sin \lambda L = 0$, and thus it is *necessary* that $\lambda = \frac{n\pi}{L}$ for some $n \in \mathbb{N}$. We thus see that each of the functions

$$u_n(t, x) = e^{-\lambda_n^2 \alpha^2 t} \sin \lambda_n x \quad \lambda_n := \frac{n\pi}{L}, \quad n \in \mathbb{N}$$

is a solution of the PDE and the boundary conditions.

Thus far, we have not taken into account the initial temperature $u(0, x) = \phi(x)$. If it should happen that $\phi(x) = B \sin \frac{n\pi x}{L}$ for some $n \in \mathbb{N}$, then the solution has now been found: It is $u(t, x) = B e^{-(\frac{n\pi\alpha}{L})^2 t} \sin \frac{n\pi x}{L}$. Of course, it is not likely that the initial condition $\phi(x)$ is of sinusoidal form. Nevertheless, we have seen that “reasonable” functions can be expanded in terms of Fourier series. Moreover, by the Principle of Superposition, a linear combination of solutions to the PDE and BC is again a solution. We thus ask if we can find constants B_n so that

$$\phi(x) = \sum_{n=1}^{\infty} B_n \sin \frac{n\pi x}{L}$$

and this is easily done:

- Extend ϕ to $[-L, 0)$ by requiring that $\phi(-x) = -\phi(x)$, and then extend ϕ to all of \mathbb{R} by requiring that it be periodic, with period $2\pi/L$.
- Note that the resulting (extended) ϕ is an odd function, and thus the Fourier expansion of ϕ will have only sine terms, and no cosine terms.
- From Fourier analysis, we know that

$$B_n = \frac{1}{L} \int_{-L}^L \phi(x) \sin \frac{n\pi x}{L} dx = \frac{2}{L} \int_0^L \phi(x) \sin \frac{n\pi x}{L} dx$$

So we have now obtained the solution:

$$u(t, x) = \sum_{n=1}^{\infty} B_n e^{-(\frac{n\pi\alpha}{L})^2 t} \sin \frac{n\pi x}{L} \quad \text{where} \quad B_n = \frac{2}{L} \int_0^L \phi(x) \sin \frac{n\pi x}{L} dx$$

□

Further Discussion:

- The separation of variables technique attempts to reduce the PDE to two ODE's. In the above example, we reduced the PDE

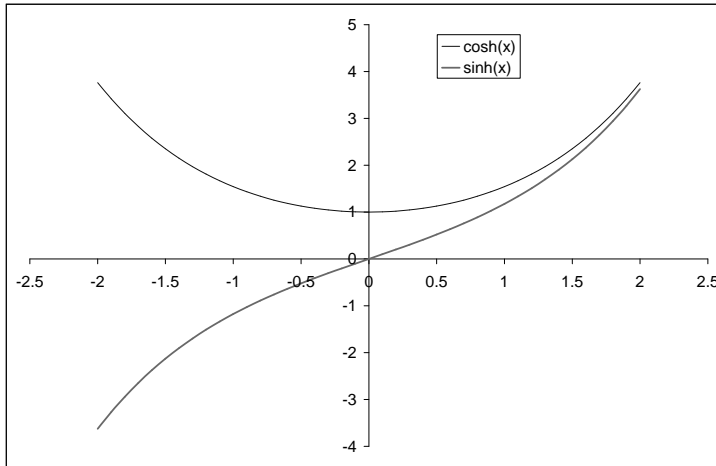
$$u_t = \alpha^2 u_{xx}$$

to the two ODE's

$$\begin{aligned} \frac{dT}{dt} &= k\alpha^2 T \\ \frac{d^2 X}{dx^2} &= kX \end{aligned}$$

Hence T is an eigenfunction (i.e. an eigenvector in a function space) of the linear operator $\frac{d}{dt}$, with eigenvalue $k\alpha^2$, and X is an eigenfunction of the linear operator $\frac{d^2}{dx^2}$, with eigenvalue k .

- Above, we glibly asserted that $k < 0$. Let us verify what happens if $k = 0$ or $k > 0$.
 - If $k = 0$, then $X(x) = A + Bx$, and $T(t)$ is constant, so we can write $u(t, x) = A + Bx$. The boundary conditions are now easily seen to imply that $A = 0 = B$.
 - If $k > 0$, we can write $k = \lambda^2$. Then $X(x) = Ae^{\lambda x} + Be^{-\lambda x}$ and $T(t) = Ce^{\lambda^2 t}$. The solution for X may also be written in terms of hyperbolic functions: Recall that $\cosh x := \frac{e^x + e^{-x}}{2}$, $\sinh x := \frac{e^x - e^{-x}}{2}$,



so that we may write $X(x) = A \cosh \lambda x + B \sinh \lambda x$ for some (different) constants A, B . We thus obtain

$$u(t, x) = e^{\lambda^2 t} [A \cosh \lambda x + B \sinh \lambda x]$$

With the help of the graphs of cosh and sinh it is easily verified that the boundary conditions imply that $A = 0 = B$ in this case also.

- The Separation of Variables technique relies on the Principle of Superposition, and thus only applies to linear PDE's. Moreover, at least one of the independent variables must be restricted to a finite interval. The domain of the problem must be consistent with the coordinate system, e.g. it must be a rectangle in cartesian coordinates, or a sector in polar coordinates, etc. The boundary conditions must be linear and homogeneous also, e.g. of the form

$$\begin{aligned} \alpha u_x(t, 0) + \beta u(t, 0) &= 0 \\ \gamma u_x(t, L) + \delta u(t, L) &= 0 \end{aligned}$$

Note, however, that there are various tricks that may be used to turn a problem with non-homogeneous boundary conditions into one with homogeneous boundary conditions.

□

Example 15.1.2 Consider a laterally insulated rod of length $L = 1$. The initial temperature is $\phi(x)$ throughout the rod. The lefthand side of the rod is kept at 0°C , whereas the righthand side is maintained at 100°C . The ICBVP for this is

$$\begin{array}{lll} \text{PDE} & u_t = \alpha^2 u_{xx} & 0 < x < 1, \quad 0 < t < \infty \\ \text{BC} & u(t, 0) = 0 \quad u(t, 1) = 100 & 0 < t < \infty \\ \text{IC} & u(0, x) = \phi(x) & \end{array}$$

If we attempt to use separation of variables, we run into trouble: Because the boundary conditions are not homogeneous, the sum of solutions will not be a solution (e.g. if u_1, u_2 are solutions, then $(u_1 + u_2)(t, 1) = 200 \neq 100$).

Nevertheless, we can employ a trick: We break up the solution u into a *steady state solution* u^S (obtained in the limit as $t \rightarrow \infty$) and a *transient solution* u^T (which will go to zero as $t \rightarrow \infty$, i.e. we write

$$u(t, x) = u^S(t, x) + u^T(t, x)$$

The *steady state solution* is, by definition, time independent, and will therefore satisfy $u_t^S = 0$. Then $u_{xx}^S = 0$ also, so $u^S(t, x) = A + Bx$. The boundary conditions are easily seen to imply that $A = 0, B = 100$, and thus the steady-state solution is $u^S(t, x) = 100x$, just as one would expect. Thus we have

$$u(t, x) = 100x + u^T(t, x) \quad \text{i.e.} \quad u^T(t, x) = u(t, x) - 100x$$

It is now easy to find an ICBVP for u^T :

$$\begin{array}{lll} \text{PDE} & u_t^T = \alpha^2 u_{xx}^T & 0 < x < 1, \quad 0 < t < \infty \\ \text{BC} & u^T(t, 0) = 0 \quad u^T(t, 1) = 0 & 0 < t < \infty \\ \text{IC} & u^T(0, x) = \phi^T(x) := \phi(x) - 100x & \end{array}$$

These boundary conditions are homogeneous. From the previous example, we now know that

$$u^T(t, x) = \sum_{n=1}^{\infty} B_n e^{-(n\pi\alpha)^2 t} \sin n\pi x \quad \text{where} \quad B_n = \frac{2}{1} \int_0^1 \phi^T(x) \sin n\pi x \, dx$$

and thus $u(t, x) = 100x + u^T(t, x)$.

□

The best way to familiarize yourself with this technique is to do the following exercises:

Exercise 15.1.3 Consider an insulated rod (laterally as well on the sides) of length L with initial temperature $\phi(x)$. In this case, the ICBVP is

$$\begin{array}{lll} \text{PDE} & u_t = u_{xx} & 0 < x < 1, \quad 0 < t < \infty \\ \text{BC} & u_x(t, 0) = 0 \quad u_x(t, 1) = 0 & 0 < t < \infty \\ \text{IC} & u(0, x) = \phi(x) & \end{array}$$

(a) Write $u(t, x) = T(t)X(x)$ and reduce the problem to two ODE's.

- (b) Deduce that $T(t) = Ce^{-\lambda^2 t}$ and that $X(x) = A \cos \lambda x + B \sin \lambda x$.
- (c) Why is $B = 0$?
- (d) Why is $\lambda = n\pi$ for some $n \in \mathbb{N}$?
- (e) We now have a solution $u_n(t, x) = A_n e^{-n^2 \pi^2 t} \cos n\pi x$ for each $n = 0, 1, 2, \dots$. Why can we say that

$$u(t, x) = \frac{A_0}{2} + \sum_{n=1}^{\infty} A_n e^{-n^2 \pi^2 t} \cos n\pi x$$

is also a solution? (We have renamed A_0 to become $\frac{A_0}{2}$ here, because this is more convenient for Fourier series.)

- (f) We now deal with the initial condition. Explain why we require that

$$\phi(x) = \frac{A_0}{2} + \sum_{n=1}^{\infty} \cos n\pi x \quad \text{for } 0 \leq x \leq 1$$

- (g) We thus need a Fourier expansion of ϕ purely in cosines. Explain how to do this. [Hint: Make ϕ an even function, and then periodic with period 2.]
- (h) Conclude that

$$A_n = 2 \int_0^1 \phi(x) \cos n\pi x \, dx \quad n = 0, 1, 2, \dots$$

- (i) What is the steady-state solution to this problem? How do you interpret this solution?

15.2 The Fundamental Solution

When we solved Laplace's equation $\Delta u = 0$, we obtained our first solution — the fundamental solution — by exploiting the an invariance property of the Laplacian operator, namely the fact that the Laplace's equation is invariant under rotations. This means that if $u(\mathbf{x})$ is a solution of Laplace's equation, and O is an orthogonal matrix, then $u(O\mathbf{x})$ is a solution also. Exercise 15.0.8 shows a similar result for the heat equation, but holds only for spatial rotations, and does not involve time. The structure of the heat equation suggests another kind of invariance. The expression $u_t - \frac{1}{2}\Delta u = 0$ contains one derivative with respect to the time variable, and two with respect to each spatial variable. Thus if $u(t, \mathbf{x})$ is a solution, so is $u(\lambda^2 t, \lambda \mathbf{x})$ (for $\lambda > 0$), i.e. if we scale space by a factor of λ and time by a factor of λ^2 , the resulting function remains a solution of the heat equation. Now if we define $r := \|\mathbf{x}\|$, then

$$\frac{(\lambda r)^2}{\lambda^2 t} = \frac{r^2}{t}$$

and this suggests that we seek a solution v which is a function of $\frac{r^2}{t}$. As in the case of Laplace's equation (where we chose a solution which is a function of just r), exploiting the symmetry of the PDE will turn it into an ODE, which we can solve.

Exercise 15.2.1 We do this for the case $n = 1$. Put $\xi := \frac{x}{\sqrt{t}}$ and suppose that $v(t, x) = v(\xi)$ is a solution of the heat equation $v_t - \frac{1}{2}v_{xx} = 0$.

$$v_t = -\frac{x}{2t^{3/2}}v' \quad v_x = \frac{1}{\sqrt{t}}v' \quad v_{xx} = \frac{1}{t}v''$$

where $v'(\xi) = \frac{dv}{d\xi}$, and obtain

$$v''(\xi) + \xi v'(\xi) = 0$$

Define $w := v'$ and integrate to obtain

$$v'(\xi) = w(\xi) = Ce^{-\xi^2/2}$$

Integrate again to obtain

$$v(\xi) = C \int_{-\infty}^{\xi} e^{-u^2/2} du$$

i.e.

$$v(t, x) = C \int_{-\infty}^{\frac{x}{\sqrt{t}}} e^{-u^2/2} du$$

□

Now that we have obtained a solution

$$u(t, x) = C \int_0^{\frac{x}{\sqrt{t}}} e^{-y^2/2} dy$$

for the one-dimensional heat equation, we may ask what initial conditions this solution satisfies. The solution is undefined at $t = 0$, so we examine the behaviour of $u(t, x)$ as $t \rightarrow 0$. Fortunately, we recognize in $e^{-y^2/2}$ a function which is closely related to the density of a standard normal random variable Z . This suggests that we choose $C = \frac{1}{\sqrt{2\pi}}$, so that

$$u(t, x) = \mathbb{P}(Z \leq \frac{x}{\sqrt{t}})$$

Now

$$\lim_{t \rightarrow \infty} \frac{x}{\sqrt{t}} = \begin{cases} \infty & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -\infty & \text{if } x < 0 \end{cases}$$

and hence

$$u(0^+, x) = \begin{cases} 1 & \text{if } x > 0 \\ \frac{1}{2} & \text{if } x = 0 \\ 0 & \text{if } x < 0 \end{cases}$$

Now note that if u solves the heat equation, so does u_x :

$$(u_x)_t = (u_t)_x = (u_{xx})_x = (u_x)_{xx}$$

i.e.

$$v_t = v_{xx} \quad \text{where } v := u_x$$

For the function u above, we have, by the Fundamental Theorem of Calculus, that $u_x(t, x) = \Phi(t, x)$ where

$$\Phi(t, x) := \frac{1}{\sqrt{2\pi t}} e^{-x^2/2t}$$

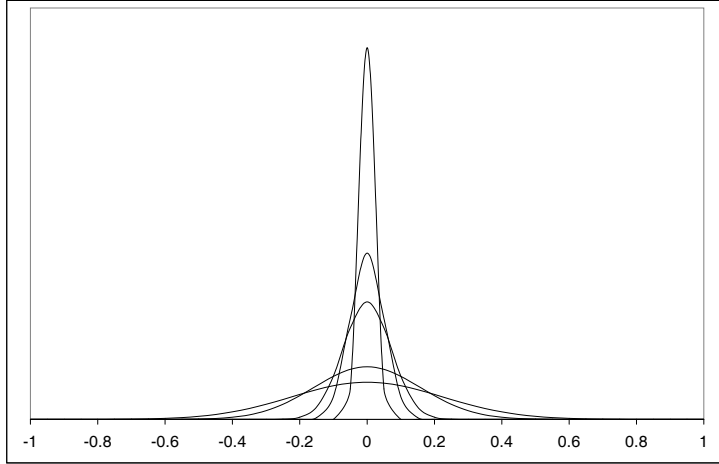
$\Phi(t, x)$ is called the *fundamental solution* of the heat equation, and also the *heat kernel*. The initial conditions corresponding to the solution Φ are quite peculiar: We see that

$$\lim_{t \rightarrow 0} \Phi(t, x) = 0 \quad \text{for all } x \neq 0$$

However, the limiting behaviour at $x = 0$ is not immediately obvious. When we remember that $\Phi(t, x)$ is the density of an $N(0, t)$ -random variable, however, we see that the graphs of the $\Phi(t, x)$ are bell curves centered at $x = 0$ that are becoming ever more peaked, in order to ensure that the total area under each curve $\Phi(t, x)$ (for fixed t) equals unity. Thus $\Phi(0, x)$ is a very peculiar function:

$$\Phi(0, x) = 0 \quad \text{for } x \neq 0, \quad \text{and} \quad \Phi(0, 0) = \infty \text{ in such a way that } \int_{-\infty}^{\infty} \Phi(0, x) dx = 1$$

Of course, if $\Phi(0, x) = 0$ for all $x \neq 0$, then $\Phi(0, x) = 0$ λ -a.e., where λ denotes Lebesgue measure, and thus (by the properties of the Lebesgue integral) $\int_{-\infty}^{\infty} \Phi(0, x) dx = \int_{-\infty}^{\infty} 0 dx = 0$. In other words, there *cannot* be a *function* $x \mapsto \Phi(0, x)$ with the required properties.



The object $\Phi(0, x)$ is therefore not a *function*. In fact, if we keep thinking probabilistically, we see that it corresponds to the distribution of a random variable with mean 0 and variance 0, i.e. the $\Phi(0, x)$ is the “density” of a random variable X which has $X = 0$ a.s., i.e. a point mass at 0. It therefore has the following important property: If $f : \mathbb{R} \rightarrow \mathbb{R}$ is a (sufficiently regular) function, then

$$\int_{-\infty}^{\infty} \Phi(0, x) f(x) dx = \mathbb{E}[f(X)] = f(0)$$

i.e. though we cannot think of $\Phi(0, x)$ as a function $\mathbb{R} \rightarrow \mathbb{R}$, we can think of it as a *functional*, i.e. a function from a set of (sufficiently regular¹) functions to \mathbb{R} :

$\Phi(0, x)$ is a rule which assigns the real number $f(0)$ to the function f

This “rule” is called the *Dirac delta function*, denoted δ_0 , and was first introduced by the physicist Paul Dirac in the 1920’s in connection with quantum mechanics. It was only in the late 1940’s that Laurent Schwartz developed a rigorous mathematical theory of such objects, called *distributions* or *generalized functions*. Though a rigorous development of generalized functions requires deep results from functional analysis, an intuitive development will provide insight into the solution of Cauchy problems, and into the nature of Green’s functions. We will tackle this soon, in the Section 15.5.

For the multidimensional case we define the fundamental solution as follows:

Definition 15.2.2 (Fundamental Solution of the Heat Equation)

The function

$$\Phi(t, \mathbf{x}) := \begin{cases} \frac{1}{(2\pi t)^{n/2}} e^{-\frac{\|\mathbf{x}\|^2}{2t}} & \mathbf{x} \in \mathbb{R}^n, t > 0 \\ 0 & \mathbf{x} \in \mathbb{R}^n, t < 0 \end{cases}$$

□

Remarks 15.2.3 Note that $\Phi(t, \mathbf{x})$ is the joint density function of a random vector (Z_1, \dots, Z_n) of independent $N(0, t)$ -variables, i.e. normal random variables with mean 0 and variance t . We therefore know that

$$\int_{\mathbb{R}^n} \Phi(t, \mathbf{x}) d\mathbf{x} = 1 \quad \text{for each } t > 0$$

Just as we did for the fundamental solution of Laplace’s equation, we will occasionally write $\Phi(t, \mathbf{x}) = \Phi(t, \|\mathbf{x}\|)$ to highlight the fact that it is radially symmetric in the spatial variable \mathbf{x} .

□

Exercise 15.2.4 (a) Verify that Φ solves the heat equation $u_t - \frac{1}{2}\Delta u = 0$.

(b) Show that Φ explodes at $(0, 0)$. in such a way that

$$\lim_{t \rightarrow 0} \int_{\mathbb{R}^n} \Phi(t, \mathbf{x}) f(\mathbf{x}) d\mathbf{x} = f(0)$$

for sufficiently regular $f : \mathbb{R}^n \rightarrow \mathbb{R}$.

[Hint: Adapt the argument for the one-dimensional $\Phi(t, x)$ given earlier.]

(c) Show that $\Phi(t, \mathbf{x}) \in C^\infty((0, \infty) \times \mathbb{R}^n)$.

□

¹“Sufficiently regular” is code for “the function is nice enough to allow me to do what I’m about to do to it”.

15.3 Solving the Heat Equation

15.3.1 The Cauchy Problem

We will first concern ourselves with the Cauchy problem on the unbounded domain $U := \mathbb{R}^+ \times \mathbb{R}^n$. Note that $\partial U = \{0\} \times \mathbb{R}^n$. We seek a function $u : \mathbb{R}^+ \times \mathbb{R}^n \rightarrow \mathbb{R}$ which solves

$$\begin{aligned} u_t - \frac{1}{2}\Delta u &= 0 && \text{in } (0, \infty) \times \mathbb{R}^n \\ u &= g \text{ on } \{0\} \times \mathbb{R}^n \end{aligned} \quad (*)$$

The “boundary condition” $u = g$ is an initial condition: $u(0, \mathbf{x}) = g(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$.

We first proceed intuitively, in the one-dimensional case: Fix a $y \in \mathbb{R}$, and regard x as a variable. Then $\Phi(t, x - y)$ is just a translation of $\Phi(t, x)$, and thus a solution of the one-dimensional heat equation. In this way, we can get, for different values $y_k \in \mathbb{R}$, many solutions $\Phi(t, x - y_k)$ of the heat equation. The heat equation is *linear*, however, and thus a *linear combination* of solutions is again a solution. In particular (regarding each y_k as fixed) the sum

$$\sum_{k=1}^m \Phi(t, x - y_k) g(y_k) \Delta y_k$$

is a solution, where $\Delta y_k := y_k - y_{k-1}$, and the “coefficients” $g(y_k) \Delta y_k$ of the linear combination are “constant”, because each y_k is regarded as fixed. In the limit, the linear combination can be made an integral:

$$u(t, x) = \int_{-\infty}^{\infty} \Phi(t, x - y) g(y) dy$$

and assuming that all functions involved are sufficiently regular to allow interchange of limit, derivative and integral (this needs proof), we guess that $u(t, x)$, as a limit of solutions to the heat equation, should itself be a solution.

To obtain the initial conditions satisfied by this solution, we adopt the probabilistic approach: For fixed t, x , the function $\Phi(t, x - y)$ is the density of a $N(x, t)$ -random variable, so

$$u(t, x) = \mathbb{E}[g(X)] \quad \text{where } X \sim N(x, t)$$

In the limit $t \rightarrow 0$, the distribution $N(x, t)$ tends to a point mass at x , and so

$$\begin{aligned} u(0, x) &= \mathbb{E}[g(X)] && \text{where } X = x \text{ a.s.} \\ &= g(x) \end{aligned}$$

We have thus shown intuitively that the function

$$u(t, x) := \int_{-\infty}^{\infty} \Phi(t, y - x) g(y) dy$$

is a solution to the Cauchy problem (*), given above. Let’s now show it formally, in n -dimensions:

Theorem 15.3.1 (Solution of Cauchy Problem for Heat Equation in \mathbb{R}^n)

Assume that $g : \mathbb{R}^n \rightarrow \mathbb{R}$ is a bounded continuous function, and define

$$u(t, \mathbf{x}) = \int_{\mathbb{R}^n} \Phi(t, \mathbf{x} - \mathbf{y}) g(\mathbf{y}) d\mathbf{y}$$

Then u is a solution of the heat equation $u_t - \frac{1}{2}\Delta u = 0$, with

$$\lim_{(t,\mathbf{x}) \rightarrow (0,\mathbf{x}_0)} u(t, \mathbf{x}) = g(\mathbf{x}_0) \quad \text{for all } \mathbf{x}_0 \in \mathbb{R}^n \quad (\text{where } t > 0, \mathbf{x} \in \mathbb{R}^n)$$

Proof: (I): We first show that u satisfies the heat equation. Since Φ blows up at $t = 0$, we fix $\delta > 0$ and consider only times $t \geq \delta$. It is straightforward to verify that the partial derivatives $\Phi_{x_i x_i}, \Phi_t$ of the fundamental solution exist and are uniformly bounded on $(\delta, \infty) \times \mathbb{R}^n$, i.e. there is a constant M_δ so that $\Phi_{x_i x_i}, \Phi_t \leq M_\delta$ for all $(t, \mathbf{x}) \in (\delta, \infty) \times \mathbb{R}^n$. It follows that for such (t, \mathbf{x}) we have

$$u_t(t, \mathbf{x}) - \frac{1}{2}\Delta u(t, \mathbf{x}) = \int_{\mathbb{R}^n} [\Phi_t(t, \mathbf{x} - \mathbf{y}) - \frac{1}{2}\Delta \Phi(t, \mathbf{x} - \mathbf{y})] g(\mathbf{y}) d\mathbf{y}$$

as interchange of integral and derivative is allowed, by the dominated convergence theorem. Since Φ is a solution to the heat equation, we see, by allowing $\delta \rightarrow 0$, that

$$u_t(t, \mathbf{x}) - \frac{1}{2}\Delta u(t, \mathbf{x}) = 0 \quad \text{for all } (t, \mathbf{x}) \in (0, \infty) \times \mathbb{R}^n$$

(II): We now tackle the initial condition. Fix $\mathbf{x}_0 \in \mathbb{R}^n$ and $\varepsilon > 0$. We want to show that $|u(t, \mathbf{x}) - g(\mathbf{x}_0)| < \varepsilon$ when (t, \mathbf{x}) lies sufficiently close to $(0, \mathbf{x}_0)$. First, by continuity of g , there is $\delta > 0$ (not the same as the δ in I.) so that

$$|g(\mathbf{y}) - g(\mathbf{x}_0)| < \varepsilon/2 \quad \text{whenever} \quad \|\mathbf{y} - \mathbf{x}_0\| < \delta$$

Suppose that $\|\mathbf{x} - \mathbf{x}_0\| < \delta/2$. Since $\int_{\mathbb{R}^n} \Phi(t, \mathbf{x} - \mathbf{y}) d\mathbf{y} = 1$, we have $g(\mathbf{x}_0) = \int_{\mathbb{R}^n} \Phi(t, \mathbf{x} - \mathbf{y}) g(\mathbf{x}_0) d\mathbf{y}$ and so

$$\begin{aligned} |u(t, \mathbf{x}) - g(\mathbf{x}_0)| &= \left| \int_{\mathbb{R}^n} \Phi(t, \mathbf{x} - \mathbf{y}) (g(\mathbf{y}) - g(\mathbf{x}_0)) d\mathbf{y} \right| \\ &\leq \int_{\mathbb{R}^n} \Phi(t, \mathbf{x} - \mathbf{y}) |g(\mathbf{y}) - g(\mathbf{x}_0)| d\mathbf{y} \\ &= \int_{B(\mathbf{x}_0, \delta)} \Phi(t, \mathbf{x} - \mathbf{y}) |g(\mathbf{y}) - g(\mathbf{x}_0)| d\mathbf{y} + \int_{\mathbb{R}^n - B(\mathbf{x}_0, \delta)} \Phi(t, \mathbf{x} - \mathbf{y}) |g(\mathbf{y}) - g(\mathbf{x}_0)| d\mathbf{y} \end{aligned}$$

We consider each of the integrals $\int_{B(\mathbf{x}_0, \delta)}, \int_{\mathbb{R}^n - B(\mathbf{x}_0, \delta)}$ in turn:

(IIa): For $\mathbf{y} \in B(\mathbf{x}_0, \delta)$ we have $|g(\mathbf{y}) - g(\mathbf{x}_0)| < \varepsilon/2$, and hence

$$\int_{B(\mathbf{x}_0, \delta)} \Phi(t, \mathbf{x} - \mathbf{y}) |g(\mathbf{y}) - g(\mathbf{x}_0)| d\mathbf{y} < \frac{\varepsilon}{2} \int_{B(\mathbf{x}_0, \delta)} \Phi(t, \mathbf{x} - \mathbf{y}) d\mathbf{y} \leq \frac{\varepsilon}{2} \int_{\mathbb{R}^n} \Phi(t, \mathbf{x} - \mathbf{y}) d\mathbf{y} = \varepsilon/2$$

(IIb): For the remaining integral, recall that g is assumed bounded, and let K be a bound, i.e. suppose that $|g(\mathbf{x})| \leq K$ for all $\mathbf{x} \in \mathbb{R}^n$. We are also still assuming that $\|\mathbf{x} - \mathbf{x}_0\| < \delta/2$. For $\mathbf{y} \in \mathbb{R}^n - B(\mathbf{x}_0, \delta)$, we have

$$\|\mathbf{y} - \mathbf{x}_0\| \leq \|\mathbf{y} - \mathbf{x}\| + \|\mathbf{x} - \mathbf{x}_0\| \leq \|\mathbf{y} - \mathbf{x}\| + \delta/2 \leq \|\mathbf{y} - \mathbf{x}\| + \frac{1}{2}\|\mathbf{y} - \mathbf{x}_0\|$$

so that

$$\|\mathbf{y} - \mathbf{x}\| \geq \frac{1}{2}\|\mathbf{y} - \mathbf{x}_0\|$$

Hence

$$\begin{aligned}
 \int_{\mathbb{R}^n - B(\mathbf{x}_0, \delta)} \Phi(t, \mathbf{x} - \mathbf{y}) |g(\mathbf{y}) - g(\mathbf{x}_0)| d\mathbf{y} &\leq 2K \int_{\mathbb{R}^n - B(\mathbf{x}_0, \delta)} \Phi(t, \mathbf{x} - \mathbf{y}) d\mathbf{y} \\
 &= \frac{2K}{(2\pi t)^{n/2}} \int_{\mathbb{R}^n - B(\mathbf{x}_0, \delta)} e^{-\|\mathbf{x} - \mathbf{y}\|^2/2t} d\mathbf{y} \\
 &\leq \frac{2K}{(2\pi t)^{n/2}} \int_{\mathbb{R}^n - B(\mathbf{x}_0, \delta)} e^{-\|\mathbf{x}_0 - \mathbf{y}\|^2/8t} d\mathbf{y} \\
 &= 2K\mathbb{P}(\|\mathbf{Z}\| \geq \delta)
 \end{aligned}$$

where $\mathbf{Z} = (Z_1, \dots, Z_n)$ and the Z_i are independent $N(\mathbf{x}_{0i}, t)$ -variables
 $\rightarrow 0$ as $t \rightarrow 0^+$

Thus for $t > 0$ sufficiently small, we can ensure that $\int_{\mathbb{R}^n - B(\mathbf{x}_0, \delta)} \Phi(t, \mathbf{x} - \mathbf{y}) |g(\mathbf{y}) - g(\mathbf{x}_0)| d\mathbf{y} < \varepsilon/2$.

(IIc): Putting (IIa) and (IIb) together, we see that $|u(t, \mathbf{x}) - g(\mathbf{x}_0)| < \varepsilon/2 + \varepsilon/2$ whenever $\|\mathbf{x} - \mathbf{x}_0\| < \delta/2$ and $t > 0$ is sufficiently small.

—

Remarks 15.3.2 We have now shown that the function $u(t, \mathbf{x}) = \int_{\mathbb{R}^n} \Phi(t, \mathbf{x} - \mathbf{y})g(\mathbf{y}) d\mathbf{y}$ is a solution to the Cauchy problem for the heat equation with initial data g on the unbounded domain $\mathbb{R}^+ \times \mathbb{R}^n$. Can there be other solutions? The answer is: yes and no. Yes, there may be other solutions to this problem. These others will all blow up, however. Using an analogue of the maximum principle (which we used to prove uniqueness of solutions to Laplace's equation) it can be shown that there is at most one solution u satisfying a growth bound

$$u(t, \mathbf{x}) \leq Ae^{B\|\mathbf{x}\|^2} \quad \text{for all } \mathbf{x} \in \mathbb{R}^n$$

for constants $A, B > 0$ i.e. all the other solutions grow extremely fast. This is *unphysical*, when dealing with actual heat problems.

It is also unfinancial: If $C(t, S)$ is the price of a call at time t before maturity, when the underlying price is S , we expect by arbitrage considerations that $C(t, S) \leq S$ — explain why! — and that $\frac{C(t, S)}{S} \rightarrow 1$ as $S \rightarrow \infty$. Assuming that C is the solution to some kind of heat equation — the topic of the next section — such rapid growth is not permitted, as it will imply that $\frac{C(t, S)}{S} \rightarrow \infty$ as $S \rightarrow \infty$, so that $C(t, S) > S$ for big enough S .

□

15.3.2 Diffusion on the Half-Line: The Method of Images

We consider the following initial-boundary value problem:

$$\begin{aligned}
 v_t - \frac{1}{2}v_{xx} &= 0 & (t, x) &\in (0, \infty) \times (0, \infty) \\
 v(0, x) &= \phi(x) & \text{for } x &\geq 0 \\
 v(t, 0) &= 0 & \text{for } t &> 0
 \end{aligned} \tag{*}$$

For intuition: This IBVP governs the evolution of the temperature of a semi-infinite bar (corresponding to the half-line $0 < x < \infty$) whose initial temperature is given by the function $g(x)$, and whose $x = 0$ -end is kept at a constant temperature of 0° .

The trick in solving this problem is to set up a Cauchy problem on the whole line $-\infty < x < \infty$, because we already know how to solve Cauchy problems on the whole line, by Theorem 15.3.1. We have to enforce initial conditions $g(x) = \phi(x)$ for $x > 0$, but we have some freedom in choosing initial conditions $g(x)$ for $x < 0$. All we need to do is to choose these initial conditions to ensure that the boundary condition $v(t, 0) = 0$ are forced to hold, by employing symmetry.

In this (very simple) case, it is easy to figure out how to do it: If we ensure that the temperature at the point $-x$ is minus the temperature at x , then the temperature at $x = 0$ must be zero:

$$u(t, -x) = -u(t, x) \implies u(t, 0) = u(t, -0) = -u(t, 0) \implies u(t, 0) = 0$$

This suggests the following Cauchy problem:

$$\begin{aligned} u_t - \frac{1}{2}u_{xx} &= 0 & (t, x) \in (0, \infty) \times (0, \infty) \\ u(0, x) &= g(x) \\ \text{where } g(x) &= \begin{cases} \phi(x) & \text{for } x \geq 0 \\ -\phi(-x) & \text{for } x < 0 \end{cases} \end{aligned}$$

Theorem 15.3.1 requires that g be continuous, but g may have a discontinuity at $x = 0$. We treat this technical point cavalierly, and write down the solution suggested by Theorem 15.3.1:

$$\begin{aligned} u(t, x) &= \int_{-\infty}^{\infty} \Phi(t, x - y)g(y) dy \\ &= \frac{1}{\sqrt{2\pi t}} \int_0^{\infty} e^{-(x-y)^2/2t} \phi(y) dy + \frac{1}{\sqrt{2\pi t}} \int_{-\infty}^0 e^{-(x-y)^2/2t} (-\phi(-y)) dy \\ &= \frac{1}{\sqrt{2\pi t}} \int_0^{\infty} \left[e^{-(x-y)^2/2t} - e^{-(x+y)^2/2t} \right] \phi(y) dy \\ &= \int_0^{\infty} [\Phi(t, x - y) - \Phi(t, x + y)] \phi(y) dy \end{aligned}$$

Thus the solution to the IBVP (*) is given by:

$$v(t, x) = \int_0^{\infty} [\Phi(t, x - y) - \Phi(t, x + y)] \phi(y) dy \quad 0 < x < \infty, 0 < t < \infty$$

where Φ is the fundamental solution of the one-dimensional heat equation.

Example 15.3.3 If we solve (*) with $\phi(x) \equiv 1$, we obtain

$$v(t, x) = \mathbb{P}(Z \geq -\frac{x}{\sqrt{t}}) - \mathbb{P}(Z \geq +\frac{x}{\sqrt{t}}) = N(\frac{x}{\sqrt{t}}) - N(-\frac{x}{\sqrt{t}})$$

where Z is a standard normal random variable, and $N(\cdot)$ the standard normal distribution function.

□

15.4 Applications to Finance

15.4.1 The Black–Scholes Option Formula

The following exercise derives the Black–Scholes option price formula by transforming the Black–Scholes PDE to a heat equation, and using the fundamental solution.

Recall the Black–Scholes PDE and boundary condition for the price of a *call option* on a share S with strike K and expiry T :

$$\begin{aligned} C_t + \frac{1}{2}\sigma^2 S^2 C_{SS} + rSC_S - rC &= 0 \\ C(T, S) &= (S - K)^+ \end{aligned} \quad (*)$$

where $C = C(t, S)$.

Exercise 15.4.1 (a) First we reduce this PDE to the one-dimensional heat equation:

(a.1) Define $x = \ln \frac{S}{K}$, $\tau = \sigma^2(T - t)$, and show that we obtain

$$\begin{aligned} v_\tau &= \frac{1}{2}v_{xx} + \gamma v_x - (\gamma + \frac{1}{2})v \\ v(0, x) &= (e^x - 1)^+ \end{aligned}$$

where $v(\tau, x) := \frac{1}{K}C(t, S)$ and $\gamma = \frac{r}{\sigma^2} - \frac{1}{2}$.

(a.2) Now define $u(\tau, x)$ by $v(\tau, x) = e^{\alpha x + \beta \tau} u(\tau, x)$. Show that

$$u_\tau = \frac{1}{2}u_{xx} + (\alpha + \gamma)u_x + \left(\frac{1}{2}\alpha^2 + \gamma\alpha - (\gamma + \frac{1}{2}) - \beta\right)u = 0$$

(a.3) Finally, show that when $\alpha := -\gamma$ and $\beta := -\frac{1}{2}(\gamma + 1)^2$, the boundary value problem (*) reduces to

$$\begin{aligned} u_\tau &= \frac{1}{2}u_{xx} \\ u(0, x) &= e^{\gamma x}(e^x - 1)^+ \end{aligned} \quad (**)$$

(b) The solution $\Phi(\tau, x)$ to the Cauchy problem

$$\begin{aligned} u_\tau &= \frac{1}{2}u_{xx} & t > 0 \\ u(x, 0) &= \delta_0 \end{aligned}$$

(where δ_0 is the Dirac delta function) is called the *fundamental solution* of the heat equation, and is given by

$$\Phi(\tau, x) = \frac{1}{\sqrt{2\pi\tau}} e^{-\frac{x^2}{2\tau}}$$

Explain why the solution of a boundary value problem

$$\begin{aligned} u_\tau &= \frac{1}{2}u_{xx} \\ u(0, x) &= f(x) \end{aligned}$$

is given by

$$u(\tau, x) = \int_{-\infty}^{\infty} \Phi(\tau, x - y) f(y) dy$$

(c) We now solve for the option price: Let $N(x) := \int_{-\infty}^x \Phi(1, y) dy$ and note that $N(x) = \mathbb{P}(Z \leq x)$, where Z is a standard normal random variable.

c.1 Use (b) to show that the solution to the BVP (**) in (a.3) is

$$u(\tau, x) = e^{(\gamma+1)x + \frac{1}{2}(\gamma+1)^2\tau} N\left(\frac{x + (\gamma+1)\tau}{\sqrt{\tau}}\right) - e^{\gamma x + \frac{1}{2}\gamma^2\tau} N\left(\frac{x + \gamma\tau}{\sqrt{\tau}}\right)$$

c.2 Finally, note that

$$C(0, S) = Kv(T, x) = Ke^{-\gamma x - \frac{1}{2}(\gamma+1)^2T} u(T, x)$$

and deduce that

$$C(0, S) = SN(d_+) - Ke^{-rT} N(d_-) \quad d_{\pm} = \frac{\ln \frac{S}{K} + (r \pm \frac{1}{2}\sigma^2)T}{\sqrt{\sigma^2 T}}$$

□

15.4.2 Barrier Options

A barrier option is a type of path-dependent option, i.e. the payoff of the option depends on the price-path of the underlying security. We will consider a *down-and-out* call option: This is like a standard (vanilla) call option C on a stock S with strike K and expiry T , except for one feature: If the underlying asset S ever reaches a barrier level $L < K$ during the life of the option, then it expires worthless. It can only have positive payoff when the minimum asset price during the life of the option is $> L$. The payoff is therefore

$$V(S, T) := (S_T - K)^+ I_{\{\min_{0 \leq t \leq T} S_t > L\}}$$

To solve this using stochastic analysis is not a trivial task. We therefore attempt to solve the appropriate boundary value problem. The relevant PDE is just the Black-Scholes PDE. But in addition to the standard boundary condition for a call option

$$V(S, T) = (S - K)^+$$

we must also take cognisance of the fact that $V = 0$ when $S = L$, i.e. that

$$V(L, t) = 0 \quad \text{for all } t \leq T$$

The appropriate BVP is therefore

$$\begin{aligned} V_t + \frac{1}{2}\sigma^2 S^2 V_{SS} + rSV_S - rV &= 0 \\ V(T, S) &= (S - K)^+ \quad \text{for } S > L \\ V(L, t) &= 0 \end{aligned} \quad (*)$$

where $V = V(t, S)$.

We first reduce the PDE to the heat equation, as in section 15.4.1: Put $x = \ln \frac{S}{K}$, $\tau = \sigma^2(T - t)$, to obtain

$$\begin{aligned} v_\tau &= \frac{1}{2}v_{xx} + \gamma v_x - (\gamma + \frac{1}{2})v \\ v(0, x) &= (e^x - 1)^+ \quad \text{for } x > \ln \frac{L}{K} \\ v(t, \ln \frac{L}{K}) &= 0 \end{aligned}$$

where $v(\tau, x) := \frac{1}{K}V(t, S)$ and $\gamma = \frac{r}{\sigma^2} - \frac{1}{2}$. Now define $u(\tau, x)$ by $v(\tau, x) = e^{\alpha x + \beta \tau} u(\tau, x)$, so that

$$u_\tau = \frac{1}{2}u_{xx} + (\alpha + \gamma)u_x + \left(\frac{1}{2}\alpha^2 + \gamma\alpha - (\gamma + \frac{1}{2}) - \beta\right)u = 0$$

Finally, when $\alpha := -\gamma$ and $\beta := -\frac{1}{2}(\gamma + 1)^2$, the boundary value problem (*) reduces to

$$\begin{aligned} u_\tau &= \frac{1}{2}u_{xx} \\ u(0, x) &= e^{\gamma x}(e^x - 1)^+ \quad \text{for } x > x_0 \\ u(t, x_0) &= 0 \\ \text{where } x_0 &:= \ln \frac{L}{K} \end{aligned} \tag{**}$$

When we solved the Black-Scholes PDE for a call option in section 15.4.1, we reduced it to a *Cauchy problem*, i.e. an initial value problem, without additional boundary conditions. Here, we have the boundary condition $u(t, x_0) = 0$, however. But we can use the *method of images* to turn this problem into an equivalent Cauchy problem, which would allow us to use the fundamental solution of the heat equation.

For intuition, consider u to be the temperature of an infinite bar. The initial condition only imposes constraints on the region of the bar where $x \geq x_0$. To ensure that the temperature is always zero at x_0 , we attempt to impose initial conditions for the region $x < x_0$ which guarantee this, i.e. which “cancel out” the temperature at x_0 . In effect, we reflect the initial data in the point x_0 . Now the reflection of the point x about x_0 is the point $\tilde{x} := x_0 - (x - x_0) = 2x_0 - x$. If we ensure that the temperature at \tilde{x} is precisely the *negative* of the temperature at x , then, by symmetry, the temperature at x_0 must be zero. We therefore want

$$u(0, x) = -u(0, \tilde{x}) = -u(0, 2x_0 - x) \quad \text{for } x \leq x_0$$

Therefore, let u be solution to the following Cauchy problem:

$$\begin{aligned} u_\tau &= \frac{1}{2}u_{xx} \\ u(0, x) &= e^{\gamma x}(e^x - 1)^+ \quad \text{for } x > x_0 \\ u(0, x) &= -e^{\gamma(2x_0 - x)}(e^{2x_0 - x} - 1)^+ \quad \text{for } x < x_0 \end{aligned} \tag{***}$$

We can write down the solution directly in terms of an integral involving the fundamental solution of the heat equation, but it turns out that a clever use of symmetry and superposition will allow us to write the solution in terms of solution obtained in section 15.4.1.

The linearity of the heat equation suggests that we break up the Cauchy problem for u into two separate problems: The first part is:

$$\begin{aligned} u_\tau^1 &= \frac{1}{2}u_{xx}^1 \\ u^1(0, x) &= e^{\gamma x}(e^x - 1)^+ \end{aligned} \tag{\dagger}$$

The other part, i.e. the Cauchy problem with antisymmetric data, is

$$\begin{aligned} u_\tau^2 &= \frac{1}{2}u_{xx}^2 \\ u^2(0, x) &= -e^{\gamma(2x_0 - x)}(e^{2x_0 - x} - 1)^+ \end{aligned} \tag{\ddagger}$$

By linearity, the sum $u^1 + u^2$ is again a solution of the heat equation. The initial conditions for the solution $u^1 + u^2$ are

$$u^1(0, x) + u^2(0, x) = u^1(0, x) - u^2(0, \tilde{x}) = e^{\gamma x}(e^x - 1)^+ - e^{\gamma(2x_0 - x)}(e^{2x_0 - x} - 1)^+$$

We examine the initial condition more closely: Note that $x_0 = \ln \frac{L}{K} < 0$, as $L < K$. Thus

$$\begin{aligned} e^{\gamma x}(e^x - 1)^+ - e^{\gamma(2x_0 - x)}(e^{2x_0 - x} - 1)^+ &= \begin{cases} e^{\gamma x}(e^x - 1)^+ & \text{if } x > 0 \\ -e^{\gamma(2x_0 - x)}(e^{2x_0 - x} - 1)^+ & \text{if } x < 2x_0 \\ 0 & \text{if } 2x_0 < x < 0 \end{cases} \\ &= \begin{cases} e^{\gamma x}(e^x - 1)^+ & \text{if } x > x_0 \\ -e^{\gamma(2x_0 - x)}(e^{2x_0 - x} - 1)^+ & \text{if } x < x_0 \end{cases} \end{aligned}$$

because $2x_0 < x_0 < 0$. Thus $u^1 + u^2$ satisfies the same initial conditions as $(***)$ and thus, by uniqueness of the solution, we see that

$$u = u^1 + u^2 \quad \text{is the solution of } (***)$$

We already know from section 15.4.1 how to solve the first part, namely

$$u^1(\tau, x) = e^{-\alpha x - \beta \tau} \frac{1}{K} C(t, S)$$

where $C(t, S)$ is the formula for the time- t price of a call option on S with strike K and expiry T .

Because the heat equation is invariant under spatial translations and rotations, we see that $u^2(\tau, x) := -u^1(\tau, \tilde{x})$ is a solution of the heat equation, because $u^1(\tau, x)$ is. Moreover, the initial conditions are $u^2(0, x) = -u^1(0, \tilde{x})$, which means that u^2 solves (\ddagger) . Thus

$$u^2(\tau, x) = -e^{-\alpha(2x_0 - x) - \beta \tau} \frac{1}{K} C(t, \frac{L^2}{S})$$

Here, the expression $\frac{L^2}{S}$ is a result of the fact that replacing x by $2x_0 - x$ is equivalent to replacing S by $\frac{L^2}{S}$: We have $S = Ke^x$, and so $Ke^{2x_0 - x} = Ke^{2 \ln \frac{L}{K} - x} = \frac{L^2}{Ke^x} = \frac{L^2}{S}$.

It follows that

$$u = e^{-\alpha x - \beta \tau} \frac{1}{K} \left[C(t, S) - e^{-2\alpha(x_0 - x)} C(t, \frac{L^2}{S}) \right]$$

Now

$$e^{-2\alpha(x_0 - x)} = e^{2\gamma(\ln \frac{L}{K} - \ln \frac{S}{K})} = \left(\frac{L}{K} \right)^{2\gamma}$$

Since the solution $V(t, S)$ of the original problem $(*)$ is just $Ke^{\alpha x + \beta \tau} u(\tau, x)$, we have

$$V(t, S) = C(t, S) - \left(\frac{L}{K} \right)^{\frac{2\gamma}{\sigma^2} - 1} C(t, \frac{L^2}{S})$$

where $C(t, S)$ is the formula for the time- t price of a call option on S with strike K and expiry T .

We have now priced a down-and-out option with strike K and barrier level $L < K$. Up-and-out options can be valued similarly. For knock-in options, *in-out parity* can be used, e.g. the equation

$$\text{Down-and-Out Call} + \text{Down-and-In Call} = \text{Vanilla Call}$$

must clearly hold by no-arbitrage.

15.5 Distributions

15.5.1 Basic Definitions

Here are some important definitions: Let $U \subseteq \mathbb{R}^n$ be an open set. Recall that the *support* of a function $\phi : U \rightarrow \mathbb{R}$ is defined by

$$\text{supp}(\phi) = \text{cl}\{x \in \mathbb{R}^n : \phi(x) \neq 0\}$$

i.e. the support of ϕ is the closure of the set where ϕ does not vanish.

Definition 15.5.1 (Test Functions)

A *test function* is a C^∞ function $\phi : U \rightarrow \mathbb{R}$ with *compact support*, i.e. a function that has partial derivatives of all orders, and which vanishes outside some compact set $K \subseteq U$.

□

An example of a test function is

$$\phi_{\mathbf{x}_0, \varepsilon}(\mathbf{x}) := \begin{cases} e^{-\frac{\varepsilon^2}{\varepsilon^2 - \|\mathbf{x} - \mathbf{x}_0\|^2}} & \text{if } \|\mathbf{x} - \mathbf{x}_0\| < \varepsilon \\ 0 & \text{else} \end{cases}$$

The function $\phi_{\mathbf{x}_0, \varepsilon}$ vanishes outside a ball of radius ε centered at \mathbf{x}_0 , and has partial derivatives of all orders.

The set of all test functions on U is denoted by $\mathcal{D}(U)$. It is easy to verify (do so!) that $\mathcal{D}(U)$ forms a linear space, with the usual (pointwise) operations of addition and scalar multiplication. We can also equip $\mathcal{D}(U)$ with a topology:

Definition 15.5.2 (Convergence of Test Functions)

If $\phi_n, \phi \in \mathcal{D}(U)$, we say that

$$\phi_n \rightarrow \phi \quad \text{in } \mathcal{D}(U)$$

if and only if

- (i) There is a compact set $K \subseteq U$ such that $\text{supp}(\phi_n) \subseteq K$ for all n (i.e. all the ϕ_n vanish outside K).
- (ii) ϕ_n and the partial derivatives of ϕ_n of arbitrary order converge to uniformly to those of ϕ .

□

Definition 15.5.3 (Distributions)

A *distribution* (or *generalized function*) is a *continuous linear map* $f : \mathcal{D}(U) \rightarrow \mathbb{R}$. This means that

$$\begin{aligned} \text{Linearity:} \quad & f(\alpha_1 \phi_1 + \alpha_2 \phi_2) = \alpha_1 f(\phi_1) + \alpha_2 f(\phi_2) \quad \phi_1, \phi_2 \in \mathcal{D}(U), \quad \alpha_1, \alpha_2 \in \mathbb{R} \\ \text{Continuity:} \quad & \text{If } \phi_n \rightarrow \phi \text{ in } \mathcal{D}(U), \text{ then } f(\phi_n) \rightarrow f(\phi) \text{ (in } \mathbb{R}) \end{aligned}$$

The set of distributions on U is denoted by $\mathcal{D}'(U)$.

□

The notation

$$\langle f, \phi \rangle := f(\phi)$$

is often used to denote the action of a distribution on a test function.

The notion of distribution generalizes that of function (hence the name *generalized function*): If $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is an ordinary Lebesgue measurable function (subject to some integrability conditions), we can consider it has a distribution by defining

$$f(\phi) := \int_{\mathbb{R}^n} f(\mathbf{x})\phi(\mathbf{x}) \, d\mathbf{x}$$

It is easy to verify that the map $f(\phi)$ is a distribution: Firstly,

$$\begin{aligned} f(\alpha_1\phi_1 + \alpha_2\phi_2) &= \int_{\mathbb{R}^n} f(\mathbf{x})(\alpha_1\phi_1(\mathbf{x}) + \alpha_2\phi_2(\mathbf{x})) \, d\mathbf{x} \\ &= \alpha_1 \int_{\mathbb{R}^n} f(\mathbf{x})\phi_1(\mathbf{x}) \, d\mathbf{x} + \alpha_2 \int_{\mathbb{R}^n} f(\mathbf{x})\phi_2(\mathbf{x}) \, d\mathbf{x} \\ &= \alpha_1 f(\phi_1) + \alpha_2 f(\phi_2) \end{aligned}$$

so $f(\phi)$ is linear. Secondly, if $\phi_n \rightarrow \phi$ in $\mathcal{D}(U)$, then the dominated convergence theorem will usually ensure that $\int_{\mathbb{R}^n} f(\mathbf{x})\phi_n(\mathbf{x}) \, d\mathbf{x} \rightarrow \int_{\mathbb{R}^n} f(\mathbf{x})\phi(\mathbf{x}) \, d\mathbf{x}$, i.e. that $f(\phi_n) \rightarrow f(\phi)$.

Not every distribution is of the form $\int_{\mathbb{R}^n} f(\mathbf{x})\phi(\mathbf{x}) \, d\mathbf{x}$, however, as we shall see shortly. Nevertheless it is customary to write

$$\int_{\mathbb{R}^n} f(\mathbf{x})\phi(\mathbf{x}) \, d\mathbf{x} \quad \text{instead of} \quad f(\phi)$$

even when f is not an ordinary function. i.e. the following expressions all mean the same thing:

$$f(\phi) = \langle f, \phi \rangle = \int_{\mathbb{R}^n} f(\mathbf{x})\phi(\mathbf{x}) \, d\mathbf{x}$$

The simplest example of a distribution which is not obtainable from an ordinary function in the above manner is:

Definition 15.5.4 (Delta Function)

The delta function is the distribution

$$\delta_0 : \mathcal{D}(\mathbb{R}^n) \rightarrow \mathbb{R} : \phi \mapsto \phi(0)$$

i.e. it is the generalized function having

$$\int_{\mathbb{R}^n} \delta_0(\mathbf{x})\phi(\mathbf{x}) \, d\mathbf{x} = \langle \delta_0, \phi \rangle = \phi(0) \quad \text{for all test functions } \phi$$

Similarly, if $\mathbf{x}_0 \in \mathbb{R}^n$, we define the delta function $\delta_{\mathbf{x}_0}$ by

$$\langle \delta_{\mathbf{x}_0}, \phi \rangle := \phi(\mathbf{x}_0)$$

i.e.

$$\int_{\mathbb{R}^n} \delta_{\mathbf{x}_0}(\mathbf{x})\phi(\mathbf{x}) \, d\mathbf{x} = \phi(\mathbf{x}_0)$$

□

Another example of a one-dimensional distribution is the map $\phi \mapsto \phi''(5)$, for example.

We recognize the delta function as the “density function” of a particular measure, the point mass at 0. This is the probability distribution whose distribution function is given by the Heaviside function:

$$H(x) := \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

If μ is an arbitrary finite measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ we can regard μ as generating a distribution (i.e. a generalized function) f_μ , as follows: For a test function ϕ , we define

$$\langle f_\mu, \phi \rangle := \int_{\mathbb{R}^n} \phi(\mathbf{x}) \mu(d\mathbf{x})$$

If μ is a probability distribution with a density f (or more generally, if $\mu \ll \lambda$ with $\frac{d\mu}{d\lambda} = f$), then

$$\int_{\mathbb{R}^n} \phi(\mathbf{x}) \mu(d\mathbf{x}) = \int_{\mathbb{R}^n} f(\mathbf{x}) \phi(\mathbf{x}) d\mathbf{x} = \langle f, \phi \rangle$$

, i.e. $D_\mu = f_\mu$ and so the distribution (generalized function) induced by μ corresponds to an ordinary function, namely its density.

For novices, distributions are best regarded as a *hybrid* of familiar objects: They are partly function, partly measure. Some of the operations one can perform on functions can be performed on distributions, but others cannot. For example, it is always possible to multiply real-valued functions, but this is not true for distributions, even in the very simplest case:

Example 15.5.5 If f, g are ordinary functions of one variable, and ϕ a test function, then

$$\langle fg, \phi \rangle = \int_{-\infty}^{\infty} (fg)\phi dx = \int_{-\infty}^{\infty} f(g\phi) dx = \langle f, g\phi \rangle$$

where the latter expression makes sense if $g\phi$ is a test function, which is the case if g is a C^∞ -function. This suggests how to define multiplication of distribution f with an ordinary C^∞ -function g :

$$\langle fg, \phi \rangle := \langle f, g\phi \rangle$$

i.e. the product of a distribution f and a smooth function g is a distribution which assigns the number $\langle f, g\phi \rangle$ to each test function ϕ . If f is an ordinary function, the the distributional product of f, g coincides with the ordinary product.

Suppose we try to make sense of the product of the delta function with itself: $\delta_0^2 = \delta_0 \times \delta_0$. Then

$$\langle \delta_0^2, \phi \rangle = \langle \delta_0, \delta_0 \phi \rangle = \delta_0(0)\phi(0)$$

and this makes no sense, because $\delta_0(0)$ makes no sense.

□

What makes distributions important for PDE theory is that we can differentiate them. If an ordinary function is differentiable, then its derivative as a distribution will coincide with its derivative as a function. However, *even non-differentiable functions become differentiable*, provided we consider them as distributions instead of ordinary functions. In particular, if distributions have partial derivatives, then they may be solutions of PDEs. We will define the derivatives of a distribution shortly. First, we need to introduce some topology on the set $\mathcal{D}'(U)$ of distributions, and for that, a notion of convergence will suffice.

15.5.2 Convergence of Distributions

Definition 15.5.6 (Convergence of distributions)

A sequence f_n in $\mathcal{D}'(U)$ converges to $f \in \mathcal{D}'(U)$ if and only if $f_n(\phi) \rightarrow f(\phi)$ for all $\phi \in \mathcal{D}(U)$, i.e.

$$\begin{aligned} f_n \rightarrow f \text{ in } \mathcal{D}'(U) &\iff \int_{\mathbb{R}^n} f_n(\mathbf{x})\phi(\mathbf{x}) \, d\mathbf{x} \rightarrow \int_{\mathbb{R}^n} f(\mathbf{x})\phi(\mathbf{x}) \, d\mathbf{x} \\ &\iff \langle f_n, \phi \rangle \rightarrow \langle f, \phi \rangle \quad \text{for all test functions } \phi \end{aligned}$$

□

Example 15.5.7 Let $\Phi(t, x) := \frac{1}{\sqrt{2\pi t}} e^{-x^2/2t}$ be the fundamental solution of the one-dimensional heat equation. In the previous section, we saw that

$$\int_{\mathbb{R}} \Phi(t, x)\phi(x) \, dx \rightarrow \phi(0) \quad \text{as } t \rightarrow 0$$

because $\int_{\mathbb{R}} \Phi(t, x)\phi(x) \, dx = \mathbb{E}[\phi(X)]$, where $X \sim N(0, t)$. Since $\phi(0) = \int_{\mathbb{R}} \delta_0(x)\phi(x) \, dx$, it follows that

$$\Phi(t, x) \rightarrow \delta_0 \quad \text{in } \mathcal{D}'(U)$$

We already know that limit $\Phi(0, x) = \lim_{t \rightarrow 0} \Phi(t, x)$ cannot be an ordinary function. We now see that it is a distribution.

□

Remarks 15.5.8 1. Convergence of distributions is inherited from convergence of measures:

It resembles one found in probability and statistics, namely convergence in distribution of random variables: If X_n, X are random variables, then we say that $X_n \rightarrow X$ in distribution if and only if the distribution functions $F_{X_n}(x)$ of the X_n converge to the distribution function $F_X(x)$ of X at every point x where F_X is continuous. It can be shown, however, that we have the following equivalence:

$$X_n \rightarrow X \text{ in distribution} \iff \mathbb{E}[\phi(X_n)] \rightarrow \mathbb{E}[\phi(X)] \quad \text{for every bounded continuous } \phi$$

If X_n, X have density functions f_n, f , then we have

$$\begin{aligned} X_n \rightarrow X \text{ in distribution} &\iff \int_{\mathbb{R}} f_n(x)\phi(x) \, dx \rightarrow \int_{\mathbb{R}} f(x)\phi(x) \, dx \\ &\text{for every bounded continuous } \phi \end{aligned}$$

and this looks very similar to the definition of convergence of distributions (i.e. of generalized functions).

2. If f_n, f are ordinary functions $\mathbb{R}^n \rightarrow \mathbb{R}$, then to say that $f_n \rightarrow f$ in distribution neither implies nor is implied by $f_n \rightarrow f$ pointwise (or λ -a.e.). For example, the functions

$$f_n(x) := nI_{(0, \frac{1}{n}]}$$

have

$$\int_{-\infty}^{\infty} f_n(x)\varphi(x) \, dx = \frac{1}{1/n} \int_0^{\frac{1}{n}} \varphi(x) \, dx = \text{average value of } \varphi \text{ on } (0, \frac{1}{n}]$$

Hence

$$\int_{-\infty}^{\infty} f_n(x) \varphi(x) dx \rightarrow \varphi(0) \text{ as } n \rightarrow \infty$$

and so $f_n \rightarrow \delta_0$ in distribution, even though $f_n \rightarrow 0$ pointwise.

3. However, if the ordinary functions f_n are dominated by an integrable function g , and if $f_n \rightarrow f$ pointwise, then the dominated convergence theorem states that

$$\int_{\mathbb{R}^n} f_n(\mathbf{x}) \varphi(\mathbf{x}) d\mathbf{x} \rightarrow \int_{\mathbb{R}^n} f(\mathbf{x}) \varphi(\mathbf{x}) d\mathbf{x}$$

and thus $f_n \rightarrow f$ in distribution as well.

□

15.5.3 Differentiation of Distributions

If $f : \mathbb{R} \rightarrow \mathbb{R}$ is an ordinary function, and ϕ a test function, then integration by parts yields

$$\int_{-\infty}^{\infty} f'(x) \phi(x) dx = f(x) \phi(x) \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} f(x) \phi'(x) dx = - \int_{-\infty}^{\infty} f(x) \phi'(x) dx$$

because $\phi(x)$ has compact support, and is therefore 0 at $\pm\infty$. (Remember that $\int_{-\infty}^{\infty}$ means $\lim_{a \rightarrow -\infty} \lim_{b \rightarrow \infty} \int_a^b$ and note that $\phi(a) = \phi(b) = 0$ for big enough a, b .) Thus we have

$$\langle f', \phi \rangle = -\langle f, \phi' \rangle$$

for ordinary functions f and test functions ϕ . This suggests that we *define* the derivative of a generalized function (distribution) f in the same way: f' is the map $f' : \mathcal{D}(U) \rightarrow \mathbb{R}$ defined by

$$f'(\phi) := -f(\phi') \quad \text{i.e.} \quad \langle f', \phi \rangle := -\langle f, \phi' \rangle$$

Note that if ϕ is a test function, then so is ϕ' , so $\langle f, \phi' \rangle$ is well defined when f is a distribution. In the multidimensional case, we adopt the same definition:

Definition 15.5.9 (Differentiation of Distributions)

Let $U \subseteq \mathbb{R}^n$ be an open set, and let $f \in \mathcal{D}'(U)$ be a distribution on U . The *partial derivative* of f w.r.t. x_j is defined as follows: $\frac{\partial f}{\partial x_j}$ is that distribution whose action on test functions ϕ is given by

$$\langle \frac{\partial f}{\partial x_j}, \phi \rangle := -\langle f, \frac{\partial \phi}{\partial x_j} \rangle$$

□

Note that

$$\langle \frac{\partial^2 f}{\partial x_j \partial x_k}, \phi \rangle = -\langle \frac{\partial f}{\partial x_k}, \frac{\partial \phi}{\partial x_j} \rangle = \langle f, \frac{\partial^2 \phi}{\partial x_j \partial x_k} \rangle$$

so the sign depends on the number of differentiations performed.

If f is an ordinary differentiable function, the integration-by-parts formula ensures that the distributional derivatives of f are the same as its ordinary derivatives. For example (taking the x_1 -coordinate for notational convenience), we have

$$\begin{aligned} \int_{\mathbb{R}^n} \frac{\partial f}{\partial x_1}(\mathbf{x}) \phi(\mathbf{x}) \, d\mathbf{x} &= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} \frac{\partial f}{\partial x_1}(x_1, \dots, x_n) \phi(x_1, \dots, x_n) \, dx_1 \right) dx_2 \dots dx_n \\ &= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \left(- \int_{-\infty}^{\infty} f(x_1, \dots, x_n) \frac{\partial \phi}{\partial x_1}(x_1, \dots, x_n) \, dx_1 \right) dx_2 \dots dx_n \\ &= - \int_{\mathbb{R}^n} f(\mathbf{x}) \frac{\partial \phi}{\partial x_1}(\mathbf{x}) \, d\mathbf{x} \end{aligned}$$

and so

$$\langle \frac{\partial f}{\partial x_1}, \phi \rangle = - \langle f, \frac{\partial \phi}{\partial x_1} \rangle$$

i.e. the ordinary partial derivative of f coincides with its distributional partial derivative. The operation of differentiation of distributions therefore extends the operation of differentiation of functions. Moreover, it is *always* defined, i.e. every distribution is differentiable, and the derivative of a distribution is another distribution. In particular, even for non-differentiable functions are differentiable, provided we allow their derivatives to be distributions.

Examples 15.5.10 1. Let's compute the (distributional) derivative of a non-differentiable function, namely the Heaviside function

$$H(x) := \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

We have

$$\langle H', \phi \rangle = - \langle H, \phi' \rangle = - \int_{-\infty}^{\infty} H(x) \phi'(x) \, dx = - \int_0^{\infty} \phi'(x) \, dx = -(\phi(\infty) - \phi(0)) = \phi(0)$$

because ϕ has compact support. Hence

$$\langle H', \phi \rangle = \langle \delta_0, \phi \rangle \quad \text{for all test functions } \phi$$

and so $H' = \delta_0$, i.e. the delta function is the distributional derivative of the Heaviside function.

2. Let's compute the derivative of the delta function:

$$\langle \delta'_0, \phi(x) \rangle = - \langle \delta_0, \phi'(x) \rangle = -\phi'(0)$$

So δ'_0 is that functional which assigns to each test function ϕ minus its derivative at zero. □

Note that distributional derivatives are very well-behaved w.r.t. convergence: If f_n, f are one-dimensional distributions, and if $f_n \rightarrow f$ in distribution, then also $f'_n \rightarrow f'$ in distribution. For if ϕ is a test function, we have

$$\lim_n \langle f'_n, \phi \rangle = - \lim_n \langle f_n, \phi' \rangle = - \langle f, \phi' \rangle = \langle f', \phi \rangle$$

where we have $\lim_n \langle f_n, \phi' \rangle = \langle f, \phi' \rangle$ because $f_n \rightarrow f$ in distribution and ϕ' is a test function. hence

$$\langle f'_n, \phi \rangle \rightarrow \langle f', \phi \rangle \quad \text{for every test function } \phi$$

and so $f'_n \rightarrow f'$ in distribution. Exactly the same argument shows that in the n -dimensional case,

$$f_n \rightarrow f \quad \text{in distribution} \quad \implies \quad \frac{\partial f_n}{\partial x_j} \rightarrow \frac{\partial f}{\partial x_j} \quad \text{in distribution}$$

15.6 Green's Functions Revisited

Since we can differentiate distributions, they may be considered as the objects in differential equations, i.e. in a PDE like $u_t - \frac{1}{2}\Delta u = f$, both u and f may be taken to be distributions. The PDE will always make sense for distributions, because they are always differentiable. Moreover, when u, f are restricted to be sufficiently smooth functions, the distributional and ordinary derivatives are the same thing. In this way, the set of possible solutions to a PDE is made bigger, and also nicer. We briefly discuss how fundamental solutions and Green's functions can be phrased much more naturally and intuitively in terms of delta functions. This is the beginning of a long and beautiful story.

15.6.1 Laplace's Equation

Consider the Laplacian operator Δ , operating on a distribution u . Since $\langle \frac{\partial^2 u}{\partial x_i^2}, \phi \rangle = \langle u, \frac{\partial^2 \phi}{\partial x_i^2} \rangle$ for any test function ϕ , we see that Δu is the distribution defined by

$$\langle \Delta u, \phi \rangle = \langle u, \Delta \phi \rangle$$

Now we saw earlier that if ϕ is a test function, then

$$\phi(0) = - \int_{\mathbb{R}^n} \Psi(\mathbf{x}) \phi(\mathbf{x}) d\mathbf{x}$$

where Ψ is the fundamental solution to Laplace's equation. Thus

$$\langle \delta_0, \phi \rangle = \phi(0) = - \langle \Psi, \Delta \phi \rangle = \langle -\Delta \Psi, \phi \rangle$$

i.e. the fundamental solution Ψ is a solution to the PDE

$$-\Delta \Psi = \delta_0$$

Distributions often provide the best way to think about fundamental solutions and Green's functions for constant coefficient PDEs.

Consider the Dirichlet problem for the Poisson equation

$$\begin{aligned} -\Delta u &= f & \text{in } U \\ u &= 0 & \text{on } \partial U \end{aligned}$$

If G is the Green's function for the region U , then we know that the solution is given by

$$u(\mathbf{x}_0) = \int_{\mathbb{R}^n} G(\mathbf{x}_0, \mathbf{x}) f(\mathbf{x}) d\mathbf{x}$$

(The surface integral vanishes because of the homogeneous boundary conditions.) Thus we have (with \mathbf{x}_0 fixed, and \mathbf{x} the variables of differentiation)

$$\langle \delta_{\mathbf{x}_0}, u \rangle = u(\mathbf{x}_0) = \langle G(\mathbf{x}_0, \mathbf{x}), f(\mathbf{x}) \rangle = \langle G^{\mathbf{x}_0}, \Delta u \rangle = \langle \Delta G^{\mathbf{x}_0}, u \rangle$$

from which it follows that, for fixed \mathbf{x}_0 , the Green's function $\mathbf{x} \mapsto G(\mathbf{x}_0, \mathbf{x})$ is the solution of the BVP

$$\begin{aligned} -\Delta G &= \delta_{\mathbf{x}_0} && \text{in } U \\ G &= 0 && \text{on } \partial U \end{aligned}$$

i.e. G is the unique distribution which solves this BVP.

15.6.2 The Heat Equation

You should already have seen that the fundamental solution Φ of the heat equation can also be related to the delta function. We repeat the argument: We know that the function

$$u(t, \mathbf{x}_0) := \int_{\mathbb{R}^n} \Phi(t, \mathbf{x} - \mathbf{x}_0) g(\mathbf{x}) d\mathbf{x}$$

is the (only nice) solution of the Cauchy problem

$$\begin{aligned} u_t - \frac{1}{2} \Delta u &= 0 \\ u(0, \mathbf{x}) &= g(\mathbf{x}) \end{aligned}$$

and that

$$u(\mathbf{x}, t) \rightarrow g(\mathbf{x}) \quad \text{as } t \rightarrow 0^+$$

Thus, with \mathbf{x}_0 fixed, it follows that

$$\langle \Phi(t, \mathbf{x}_0 - \mathbf{x}), g \rangle = u(t, \mathbf{x}_0) \rightarrow g(\mathbf{x}_0) = \langle \delta_{\mathbf{x}_0}, g \rangle$$

for all g and thus all test functions. In particular, with $\mathbf{x}_0 = 0$, and noting that $\Phi(t, \mathbf{x}) = \Phi(t, -\mathbf{x})$, we have

$$\langle \Phi(t, \mathbf{x}), g \rangle \rightarrow \langle \delta_0, g \rangle \quad \text{as } t \rightarrow 0^+$$

and thus that

$$\langle \Phi(t, \mathbf{x}) \rightarrow \delta_0(\mathbf{x}) \quad \text{in distribution}$$

So the fundamental solution Φ is the solution to the BVP

$$\begin{aligned} \Phi_t - \frac{1}{2} \Delta \Phi &= 0 && \text{in } (0, \infty) \times \mathbb{R}^n \\ \Phi(0, \mathbf{x}) &= \delta_0 \end{aligned}$$

Chapter 16

The Radon–Nikodým Theorem*

16.1 Definitions and Statement of Radon–Nikodým Theorem

We begin with some definitions:

Definition 16.1.1 Let ν, μ be measures on a measurable space (S, \mathcal{S}) .

- (i) We say that ν is *absolutely continuous* w.r.t. μ , and write $\nu \ll \mu$, iff $\mu A = 0$ implies $\nu A = 0$ for all $A \in \mathcal{S}$.
- (ii) μ, ν are *equivalent* iff $\nu \ll \mu$ and $\mu \ll \nu$.
- (iii) We say that μ, ν are *mutually singular*, and write $\mu \perp \nu$, iff there exists $A \in \mathcal{S}$ such that $\mu A^c = \nu A = 0$.

□

Remarks 16.1.2 (a) Two measures are equivalent iff they have the same null sets.

(b) Two probability measures are equivalent iff they have the same null sets, iff they have the same sets of measure 1, iff they have the same sets of positive measure.

(c) Two measures are mutually singular iff their “masses” are concentrated on disjoint sets: If $\mu A = \nu A^c = 0$, then all the mass of μ lies in A^c , and all the mass of ν lies in A .

□

Examples 16.1.3 (a) Suppose that (S, \mathcal{S}, μ) is a measure space, and that $f \in \mathcal{S}^+$. Recall from Propn. 6.4.1 that there is a measure $\nu = f \cdot \mu$ on (S, \mathcal{S}) , defined by

$$\nu A = \int_A f \, d\mu$$

It is clear that $\nu \ll \mu$.

The map f is called the *density*, or *Radon–Nikodým derivative*, of ν w.r.t. μ , and also denoted $\frac{d\nu}{d\mu}$.

The Radon–Nikodým Theorem (below) states that the above way of constructing an absolutely continuous measure is the *only* way to do so: If $\nu \ll \mu$, then ν has a density, i.e. then $\nu = f \cdot \mu$ for some non-negative measurable f .

Also recall that if $\nu = f \cdot \mu$, then $\nu g = \mu(fg)$ (cf. Propn. 6.4.2, the Chain Rule).

(b) Clearly any point mass is singular w.r.t. Lebesgue measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, e.g. $\delta_0 \perp \lambda$.

□

Here is the main result of this section:

Theorem 16.1.4 (Radon–Nikodým)

Suppose that ν, μ are σ -finite measures on (S, \mathcal{S}) and that $\nu \ll \mu$. Then there exists a μ -a.e. unique $f \in m\mathcal{S}^+$ such that $\nu = f \cdot \mu$.

□

The next subsection is devoted to its proof, which involves a number of interesting related results.

We have the following versions of the Chain Rule and Reciprocal Rule for densities:

Proposition 16.1.5 (a) Suppose that ν, η, μ are measures on a common measurable space, and that $\nu \ll \eta$ and $\eta \ll \mu$. Then also $\nu \ll \mu$ and

$$\frac{d\nu}{d\mu} = \frac{d\nu}{d\eta} \frac{d\eta}{d\mu} \quad \mu\text{-a.e.}$$

$$\text{i.e. } f \cdot (g \cdot \mu) = (fg) \cdot \mu$$

(b) If $\nu \ll \mu$ and $\frac{d\nu}{d\mu} > 0$ μ -a.e., then also $\mu \ll \nu$, and

$$\frac{d\mu}{d\nu} = \left(\frac{d\nu}{d\mu} \right)^{-1} \quad \mu\text{-a.e.}$$

$$\text{i.e. if } \nu = f \cdot \mu \text{ and } f > 0, \text{ then } \mu = \frac{1}{f} \cdot \nu.$$

□

Exercise 16.1.6 Prove Propn. 16.1.5.

□

16.2 Proof of the Radon–Nikodým Theorem; Related Results

The proof of the Radon–Nikodým Theorem involves a number of intermediate steps, all interesting in their own right: We will prove the Hahn–, Jordan– and Lebesgue Decomposition Theorems in succession. The Radon–Nikodým Theorem is a special case of the latter.

First, we generalize the concept of measure:

Definition 16.2.1 Let (S, \mathcal{S}) be a measurable space. A *signed measure* is a countably additive map $\alpha : \mathcal{S} \rightarrow [-\infty, \infty]$ such that $\alpha\emptyset = 0$.

α is said to be *finite* if $-\infty < \alpha A < +\infty$ for all $A \in \mathcal{S}$.

A set $P \subseteq S$ is called *positive* for α iff $\alpha A \geq 0$ for every measurable $A \subseteq P$.

Negative sets are defined analogously.

□

Remarks 16.2.2 (a) Instead of length, volume or mass, think of αA as the electric *charge* of the set A .

(b) Note that \emptyset is both positive and negative.

(c) Note that a signed measure α on (S, \mathcal{S}) cannot take on *both* the values $\pm\infty$. Indeed, if $A \in \mathcal{S}$, then the sum $\alpha A + \alpha A^c$ must be defined, and must be equal to αS . Sums of the form $\infty - \infty$ and $(-\infty) + (+\infty)$ are not allowed. Hence if there is some A for which $\alpha A = +\infty$, then $\alpha S = +\infty$, and if there is some B for which $\alpha B = -\infty$, then $\alpha B = -\infty$. Hence we cannot find measurable A, B such that $\alpha A = +\infty, \alpha B = -\infty$.

□

Exercise 16.2.3 Show that the union of countably many positive (negative) sets is positive (negative).

□

Exercise 16.2.4 Suppose that α is a signed measure on (S, \mathcal{S}) . Suppose that $A_n \in \mathcal{S}$, for $n \in \mathbb{N}$, and that either that

(i) $A_n \uparrow A$, or

(ii) $A_n \downarrow A$, and $-\infty < \alpha A_n < \infty$ for some $n \in \mathbb{N}$.

Prove that $\alpha A_n \rightarrow \alpha A$.

[Hint: see Exercise 4.3.5.]

□

Theorem 16.2.5 (Hahn Decomposition Theorem)

Suppose that α is a signed measure on the measurable space (S, \mathcal{S}) . Then there exists a positive measurable set S^+ and a negative measurable set S^- in \mathcal{S} such that

$$S^+ \cup S^- = S \quad S^+ \cap S^- = \emptyset$$

Before we can prove this result, we need a definition and a lemma:

Lemma 16.2.6 Let α be a signed measure on (S, \mathcal{S}) , and suppose $A \in \mathcal{S}$ satisfies $-\infty < \alpha A < 0$. Then there is a negative set $B \in \mathcal{S}$ such that $B \subseteq A$ and $\alpha B \leq \alpha A$.

Proof: We construct a sequence δ_n of non-negative (extended) reals and a sequence A_n of subsets of A as follows: Let

$$\delta_1 := \sup\{\alpha E : E \in \mathcal{S}, E \subseteq A\}$$

Then $\delta_1 \geq \alpha \emptyset = 0$. Choose $A_1 \in \mathcal{S}$ such that $A_1 \subseteq A$ and $\alpha A_1 \geq \frac{\delta_1}{2} \wedge 1$ (where we have taken the \vee with 1 because it's not yet obvious that $\delta_1 < \infty$).

Given A_1, \dots, A_{n-1} , define

$$\delta_n := \sup\left\{\alpha E : E \in \mathcal{S}, E \subseteq A - \bigcup_{k=1}^{n-1} A_k\right\}$$

(so that $\delta_n \geq 0$), and then choose a measurable $A_n \subseteq A - \bigcup_{k=1}^{n-1} A_k$ such that $\alpha A_n \geq \frac{\delta_n}{2} \wedge 1$. Now put

$$A_\infty = \bigcup_n A_n \quad B = A - A_\infty$$

Note that $\alpha A_\infty = \sum_n \alpha A_n \geq 0$ (because the A_n are mutually disjoint), and $\alpha A = \alpha A_\infty + \alpha B \geq \alpha B$.

It therefore remains to show that B is a negative set. First note that $\delta_n \rightarrow 0$: For

$$\alpha A = \alpha A_\infty + \alpha B \quad \alpha A \text{ is finite}$$

together imply that αA_∞ is finite, and since $\alpha A_\infty = \sum_n \alpha A_n$, we must have $\alpha A_n \rightarrow 0$. Since $0 \leq \frac{\delta_n}{2} \wedge 1 \leq \alpha A_n$, we must have $\delta_n \rightarrow 0$ as well.

Now if E is a measurable subset of $B = A - A_\infty$, then E is a subset of each $A - \bigcup_{k=1}^{n-1} A_k$ as well, and hence $\alpha E \leq \delta_n$, for all $n \in \mathbb{N}$. Thus $\alpha E \leq 0$, proving that B is negative.

+

Proof of Hahn Decomposition Theorem: Since $+\infty, -\infty$ cannot both be amongst the values of α , assume that $-\infty < \alpha A < \infty$ for every $A \in \mathcal{S}$.

$$a = \inf\{\alpha N : N \text{ is negative}\}$$

Then $a \leq \alpha \emptyset = 0$. Let N_n be a sequence of positive sets such that $\alpha N_n \downarrow a$. The union of negative sets is negative, so we may assume that N_n is increasing. Let $S^- := \bigcup_n N_n$. By Exercise 16.2.3, S^- is a negative set. By Exercise 16.2.4, $\alpha S^- = \lim_{n \rightarrow \infty} \alpha N_n = a$, and thus a is finite.

Let $S^+ := S - S^-$. It remains to check that S^+ is positive. But if $A \subseteq S^+$ is a measurable set with $\alpha A < 0$, then there is a negative set $B \subseteq A$ such that $\alpha B \leq \alpha A < 0$. Now clearly $S^- \cap B = \emptyset$, and so $S^- \cup B$ is a negative set with $\alpha(S^- \cup B) = \alpha S^- + \alpha B < \alpha S^- = a$, contradicting the definition of a .

+

An immediate corollary is the following:

Theorem 16.2.7 (Jordan Decomposition Theorem)

Every signed measure is the difference of two mutually singular (non-negative) measures, at least one of which is finite.

Proof: Let S^+, S^- be the subsets of S given by the Hahn Decomposition Theorem. Define two maps $\alpha^\pm : \mathcal{S} \rightarrow [0, \infty]$ by

$$\alpha^+ A = \alpha(A \cap S^+) \quad \alpha^- A = -\alpha(A \cap S^-)$$

Then clearly, α^+, α^- are non-negative measures, and $\alpha = \alpha^+ - \alpha^-$. Now at least one of $\alpha S^+, \alpha S^-$ is finite, and hence at least one of α^+, α^- is a finite measure.

+

Theorem 16.2.8 (Lebesgue Decomposition)

For σ -finite measures μ, ν on a measurable space (S, \mathcal{S}) , there exist unique measures ν_a, ν_c on (S, \mathcal{S}) such that $\nu_a \ll \mu, \nu_s \perp \mu$ and $\nu = \nu_a + \nu_s$. Moreover, $\nu_a = f \cdot \mu$ for some μ -a.e. unique $f \in m\mathcal{S}^+$.

Before we can prove this theorem, we need two lemmas:

Lemma 16.2.9 On a measurable space (S, \mathcal{S}) , suppose that we have two measures μ, ν , and a sequence of measurable functions $f_n \in m\mathcal{S}^+$ such that $f_n \cdot \mu \leq \nu$ (for $n \in \mathbb{N}$). Let $f = \sup_n f_n$. Then also $f \cdot \mu \leq \nu$.

Proof: Suppose that $f, g \in m\mathcal{S}^+$ with $f \cdot \mu, g \cdot \mu \leq \nu$, and define $h := f \vee g, A := \{f \geq g\}$. If $B \in \mathcal{S}$, then

$$(h \cdot \mu)B = \int_B h \, d\mu = \int_{A \cap B} f \, d\mu + \int_{A^c \cap B} g \, d\mu \leq \nu(A \cap B) + \nu(A^c \cap B) = \nu B$$

and hence $h \cdot \mu \leq \nu$.

Now, given the sequence f_n , define $g_n := f_1 \vee \cdots \vee f_n$ (for $n \in \mathbb{N}$). Then $g_n \cdot \mu \leq \nu$, and $g_n \uparrow f$. By the MCT, $g_n \cdot \mu \uparrow f \cdot \mu$, and hence $f \cdot \mu \leq \nu$ also.

+

Lemma 16.2.10 Let μ, ν be finite (non-negative) measures on a measurable space (S, \mathcal{S}) , such that $\nu \not\ll \mu$. Then there exists $f \in m\mathcal{S}^+$ such that $\mu f > 0$ and $f \cdot \mu \leq \nu$.

Proof: For $n \in \mathbb{N}$, define a signed measure $\gamma_n := \nu - n^{-1}\mu$, and let S_n^\pm be two sets given by the Hahn Decomposition Theorem applied to γ_n . Since $\gamma_1 \leq \gamma_2 \leq \cdots$, we may assume that $S_1^+ \subseteq S_2^+ \subseteq \cdots$, because if $n < m$, then S_n^+ is positive for γ_m . Let $A := \bigcup_n S_n^+$, and so that $A^c = \bigcap_n S_n^-$. Then for each $n \in \mathbb{N}$,

$$0 \leq \nu A^c \leq \nu S_n^- = \gamma_n S_n^- + n^{-1}\mu S_n^- \leq n^{-1}\mu S$$

because $\gamma_n S_n^- \leq 0$. Now $n^{-1}\mu S \rightarrow 0$ as $n \rightarrow \infty$, and hence $\nu A^c = 0$. Since $\mu \not\ll \nu$, we must have $\mu A > 0$. It follows that $\mu S_n^+ > 0$ for some n , because $S_n^+ \uparrow A$. Define $f := n^{-1}I_{S_n^+}$, so that $\mu f > 0$. If $B \in \mathcal{S}$, then

$$(f \cdot \mu)B = n^{-1}\mu(S_n^+ \cap B) = \nu(S_n^+ \cap B) - \gamma_n(S_n^+ \cap B) \leq \nu B$$

because S_n^+ is positive for γ_n , and hence $f \cdot \mu \leq \nu$.

+

Proof of the Lebesgue Decomposition Theorem: First assume that μ, ν are finite measures. Let

$$\mathcal{C} := \{f \in m\mathcal{S}^+ : f \cdot \mu \leq \nu\} \quad c := \sup\{\mu f : f \in \mathcal{C}\}$$

Choose $f_n \in \mathcal{C}$ so that $\mu f_n \rightarrow c$, and let $f := \sup_n f_n$. As in Lemma 16.2.9, we may assume $f_n \uparrow f$, so that $\mu f_n \uparrow \mu f$ (by MCT), and hence $\nu f = c$. Lemma 16.2.9 implies that $f \in \mathcal{C}$.

Define $\nu_a := f \cdot \mu$, and, perforce, $\nu_s := \nu - \nu_a$. It is clear that $\nu_a \ll \mu$, so we must check that $\nu_s \perp \mu$.

Suppose not. Invoke Lemma 16.2.10 to obtain a measurable $g \geq 0$ with $\mu g > 0$ and $g \cdot \mu \leq \nu_s$. Then $f + g \in \mathcal{C}$, because $(f + g) \cdot \mu = \nu_a + g \cdot \mu \leq \nu_a + \nu_s$, and $\mu(f + g) > c$ — contradiction. Hence $\nu_s \perp \mu$.

We now have a decomposition $\nu = \nu_a + \nu_s$. To prove that it is *unique*, assume that we have another decomposition $\nu = \bar{\nu}_a + \bar{\nu}_s$, with $\bar{\nu}_a \ll \mu$, $\bar{\nu}_s \perp \mu$. Choose $A, \bar{A} \in \mathcal{S}$ such that

$$\nu_s A = \mu A^c = 0 \quad \bar{\nu}_s \bar{A} = \mu \bar{A}^c = 0$$

and let $B = A \cap \bar{A}$. Then clearly $\nu_s(B) = 0$. Now $0 \leq \mu(A^c \cup \bar{A}^c) \leq \mu A^c + \mu \bar{A}^c = 0$, so that also $\nu_a(B^c) = 0$. It follows that, for $C \in \mathcal{S}$, we have $\nu_a C = \nu_a(B \cap C)$, and thus

$$(I_B \cdot \nu)C = \nu(B \cap C) = \nu_a(B \cap C) = \nu_a C$$

We have therefore shown that

$$\nu_a = I_B \cdot \nu$$

Similarly $\bar{\nu}_s(B) = 0$, $\bar{\nu}_a(B^c) = 0$ imply that $\bar{\nu}_a = I_B \cdot \nu$. Hence $\nu_a = \bar{\nu}_a$, and so $\nu_s = \nu - \nu_a = \nu - \bar{\nu}_a = \bar{\nu}_s$.

Next, we must check that the f in $\nu_a := f \cdot \mu$ is unique μ -a.e. Suppose that also $\nu_a = g \cdot \mu$ for some $g \in m\mathcal{S}^+$, and let $h = f - g$. Then $h \cdot \mu = 0$, and hence

$$\mu|h| = \int_{\{h>0\}} h \, d\mu - \int_{\{h<0\}} h \, d\mu = 0$$

so that $h = 0$ μ -a.e. by Lemma 6.3.8.

It remains to drop the assumption that μ, ν are finite measures. If they are merely σ -finite, we can choose a sequence of mutually disjoint measurable sets A_n with union $\bigcup_n A_n = S$ such that $\mu A_n, \nu A_n < \infty$ for all $n \in \mathbb{N}$. Applying the above to the measurable space $(A_n, \mathcal{S} \cap A_n)$, we are able to find a μ -a.e. unique $\mathcal{S} \cap A_n$ -measurable $f_n : A_n \rightarrow \mathbb{R}$ and a unique ν_s^n such that $\nu = f_n \cdot \mu + \nu_s^n$ and $\nu_s^n \perp \mu$ on the measurable space $(A_n, \mathcal{S} \cap A_n)$. Define $f : S \rightarrow \mathbb{R}$ by gluing the f_n 's:

$$f(s) := f_n(s) \quad \text{iff} \quad s \in A_n$$

It is easy to see that $f \in m\mathcal{S}^+$. Define $\nu_s := \nu - f \cdot \mu$, and note that if $C \subseteq A_n$, then

$$\nu_s C = \nu C - (f \cdot \mu)C = \nu C - (f_n \cdot \mu)C = \nu_s^n C$$

To see that $\nu_s \perp \mu$ on (S, \mathcal{S}) is now straightforward: Choose $B_n \subseteq A_n$ are such that $\nu_s^n B_n = 0$, $\mu(A_n - B_n) = 0$ (which exist because $\nu_s^n \perp \mu$ on (A_n, \mathcal{S}_n)), and let $B := \bigcup_n B_n$. Then

$$\nu_s B = \sum_n \nu_s B_n = \sum_n \nu_s^n B_n = 0$$

and, using the disjointness of the A_n and the fact that $B_n \subseteq A_n$,

$$\mu B^c = \mu \left(\bigcup_n (A_n - B_n) \right) = \sum_n \mu(A_n - B_n) = 0$$

As a corollary, we have

Theorem 16.2.11 (Radon–Nikodým)

Suppose that ν, μ are σ -finite measures on (S, \mathcal{S}) and that $\nu \ll \mu$. Then there exists a μ -a.e. unique $f \in m\mathcal{S}^+$ such that $\nu = f \cdot \mu$.

□

16.3 Products

16.3.1 Introduction

Example 16.3.1 (a) Denote by μA the *area* of a subset A of \mathbb{R}^2 . We know how to define μ on *rectangles*, i.e. sets of the form $A = B_1 \times B_2$, where B_1, B_2 are intervals in \mathbb{R} : Indeed

$$\mu A = \lambda B_1 \times \lambda B_2 \quad (*)$$

where λ is Lebesgue measure. So μ is to be a measure on $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ such that $\mu(B_1 \times B_2) = \lambda(B_1)\lambda(B_2)$. Of course, many sets in $\mathcal{B}(\mathbb{R}^2)$ do not have the form $B_1 \times B_2$, and we would like μ to be defined for them as well. So $(*)$ cannot serve as a definition of μ .

- (b) In probability theory, it is quite natural to consider the product of two probability spaces. Such products typically model sequences of independent experiments. For example, let $\Omega_1 = \{H, T\}$, $\mathcal{F}_1 = \mathcal{P}(\Omega_1)$ and let $\mathbb{P}_1\{H\} = \frac{1}{2} = \mathbb{P}_1\{T\}$. Then $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ models the tossing of a fair coin. Now let $\Omega_2 = \{1, 2, \dots, 6\}$, $\mathcal{F}_2 = \mathcal{P}(\Omega_2)$ and $\mathbb{P}_2\{1\} = \mathbb{P}_2\{2\} = \dots = \mathbb{P}_2\{6\} = \frac{1}{6}$. Then $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ models the rolling of a fair die. The underlying set of the probability space which models the *combined* random experiment “First toss a fair coin, and then roll a fair die” can clearly be taken to be the cartesian product $\Omega = \Omega_1 \times \Omega_2$. The natural σ -algebra will be $\mathcal{F} = \mathcal{P}(\Omega_1 \times \Omega_2)$, and it is not hard to see that this σ -algebra is generated by the π -system $\{B_1 \times B_2 : B_1 \in \mathcal{F}_1, B_2 \in \mathcal{F}_2\}$. Now the event $B_1 \times B_2 \subseteq \Omega$ consists of all those outcomes $\omega = (\omega_1, \omega_2) \in \Omega_1 \times \Omega_2$ having $\omega_1 \in B_1$ and $\omega_2 \in B_2$. Thus $B_1 \times B_2$ occurs in the combined random experiment iff B_1 and B_2 occur in each of the individual experiments.

The probability measure associated with the combined random experiment would therefore naturally satisfy

$$\mathbb{P}(B_1 \times B_2) = \mathbb{P}_1(B_1)\mathbb{P}_2(B_2) \quad (**)$$

But not every event in $\mathcal{P}(\Omega_1 \times \Omega_2)$ is of the form $B_1 \times B_2$, so $(**)$ cannot serve as a definition of \mathbb{P} .

□

The aim of this section is to construct, out of two measure spaces $(S, \mathcal{S}, \mu), (T, \mathcal{T}, \nu)$ an new measure space $(S \times T, \mathcal{S} \otimes \mathcal{T}, \mu \otimes \nu)$ satisfying the following requirements:

- (i) A subset of $S \times T$ is called a *measurable rectangle* if it has the form $A \times B$, where $A \in \mathcal{S}, B \in \mathcal{T}$.
 $\mathcal{S} \otimes \mathcal{T}$ is *defined* to be the smallest σ -algebra on $S \times T$ which has all rectangles with measurable sides as members.

(ii) For each *rectangle* $A \times B$, we require that $(\mu \otimes \nu)(A \times B) = \mu A \cdot \nu B$

Remarks 16.3.2 (a) A remark on notation: We will be working with functions of more than one variable, and may integrate with respect to just one of those variables. We therefore introduce the following notation:

$$\int f(x) \mu(dx) := \mu f =: \mu^x f(x)$$

Thus, for example, $\int f(x, y) \mu(dx)$ integrates the function $f(x, y)$ over x , keeping y fixed. The integral $\iint f(x, y) \mu(dx) \nu(dy)$ is a double integral that first integrates f w.r.t. μ over the variable x , and then integrates the function $y \mapsto \int f(x, y) \mu(dx)$ w.r.t. ν over the variable y . We may also write this as $\nu^y(\mu^x f(x, y))$.

(b) Several times below, we will prove a result for *finite* measures, and then refer to a “standard argument” to lift the result to σ -finite measures. This is done as follows: Suppose that μ is σ -finite on (S, \mathcal{S}) , and that a result Φ has been proved to hold for finite measures. Since μ is σ -finite, there exists a sequence of measurable sets $A_n \uparrow S$ such that $\mu A_n < \infty$ for all $n \in \mathbb{N}$. The measures $\mu_n := I_{A_n} \cdot \mu$ are *finite* on (S, \mathcal{S}) , so that result Π holds for the μ_n . By the MCT, if $f \in m\mathcal{S}^+$, then

$$\mu f = \mu(\lim_n f I_{A_n}) = \lim_n \mu_n f$$

This is often enough to show that Φ holds for μ as well.

□

16.3.2 Products of Measure Spaces

Given two measurable spaces $(S, \mathcal{S}), (T, \mathcal{T})$, there we can construct a σ -algebra $\mathcal{S} \otimes \mathcal{T}$ on the cartesian product $S \times T$:

Definition 16.3.3 Let (S, \mathcal{S}) and (T, \mathcal{T}) be measurable spaces. Define *projections* $\pi_S : S \times T \rightarrow S$, $\pi_T : S \times T \rightarrow T$ by

$$\pi_S : (s, t) \mapsto s \qquad \pi_T : (s, t) \mapsto t$$

Then define $\mathcal{S} \otimes \mathcal{T} := \sigma(\pi_S, \pi_T)$ to be the smallest σ -algebra for which both projections are measurable.

□

Exercise 16.3.4 Let (S, \mathcal{S}) and (T, \mathcal{T}) be measurable spaces, and let $\mathcal{R} := \{A \times B : A \in \mathcal{S}, B \in \mathcal{T}\}$ be the set of all measurable rectangles. Note that \mathcal{R} is a π -system. Show that $\mathcal{S} \otimes \mathcal{T} = \sigma(\mathcal{R})$.

Hence the product σ -algebra is generated by the π -system of all measurable rectangles.

[Hint: $A \times B = (A \times T) \cap (S \times B)$, and $A \times T = \pi_S^{-1}[A]$.]

□

Exercise 16.3.5 Show that $\mathcal{B}(\mathbb{R}^2) = \mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R})$.

[Hint: Using Exercise 16.3.4, it is easy to see that $\mathcal{B}(\mathbb{R}^2) \supseteq \mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R})$. For the opposite direction, show that any open set in \mathbb{R}^2 can be written as a countable union of sets of the form $U \times V$, where U, V are open intervals in \mathbb{R} .]

□

Suppose that (S, \mathcal{S}, μ) and (T, \mathcal{T}, ν) are measure spaces. We would like to construct a measure $\mu \otimes \nu$ on $(S \times T, \mathcal{S} \otimes \mathcal{T})$. One way that suggests itself is to define

$$(1) \quad (\mu \otimes \nu)B := \int \left(\int I_B(s, t) \nu(dt) \right) \mu(ds) = \mu^s(\nu^t I_B(s, t)) \quad B \in \mathcal{S} \otimes \mathcal{T}$$

Another is to define it as

$$(2) \quad (\mu \otimes \nu)B := \int \left(\int I_B(s, t) \mu(ds) \right) \nu(dt) = \nu^t(\mu^s I_B(s, t)) \quad B \in \mathcal{S} \otimes \mathcal{T}$$

Exercise 16.3.6 Check that

$$(\mu \otimes \nu)(A \times B) = \mu A \cdot \nu B \quad A \in \mathcal{S}, B \in \mathcal{T}$$

for both of the above possible definitions of $\mu \otimes \nu$.

□

We shall soon see that (i) the above definitions are both possible, and (ii) they coincide.

We first investigate the possibility of defining $\mu \otimes \nu$ in the above manner. To be able to do perform a double integral $\iint f(s, t) \nu(dt) \mu(ds)$ it is necessary that:

- (i) for each $s \in S$, the map $t \mapsto f(s, t)$ must \mathcal{T} -measurable, so that we can calculate the inner integral $\int f(s, t) \nu(dt)$;
- (ii) the map $s \mapsto F(s) := \int f(s, t) \nu(dt)$ must be \mathcal{S} -measurable, so that we can calculate the outer integral $\int F(s) \mu(ds)$.

The following lemma gives us what we need:

Lemma 16.3.7 Suppose that (S, \mathcal{S}) and (T, \mathcal{T}) are measurable spaces, that μ is a σ -finite measure on (S, \mathcal{S}) , and that $f : S \times T \rightarrow \mathbb{R}^+$ is $\mathcal{S} \otimes \mathcal{T}$ -measurable. Then

- (i) For each $t \in T$, the map $s \mapsto f(s, t)$ is \mathcal{S} -measurable.
- (ii) The map $t \mapsto \int f(s, t) \mu(ds)$ is \mathcal{T} -measurable.

Proof: We apply the Monotone Class Theorem (Thm. 8.1.5. First assume that μ is a finite measure, and let

$$\mathcal{H} = \{f \in m\mathcal{S} \otimes \mathcal{T} : f \text{ is bounded and satisfies (i) and (ii)}\}$$

It is easy to verify that \mathcal{H} is a vector space (we need the finiteness of μ in order to avoid expressions of the form $\infty - \infty$), and that that each $I_{A \times B} \in \mathcal{H}$, where $A \in \mathcal{S}, B \in \mathcal{T}$. By the MCT, \mathcal{H} is closed under bounded limits of increasing non-negative sequences. Moreover, the set $\mathcal{R} := \{A \times B : A \in \mathcal{S}, B \in \mathcal{T}\}$ is a π -system with the property that $I_R \in \mathcal{H}$ for

every $R \in \mathcal{R}$, and thus by Thm. 8.1.5 every bounded $\mathcal{S} \otimes \mathcal{T}$ -measurable function belongs to \mathcal{H} (since $\sigma(\mathcal{R}) = \mathcal{S} \otimes \mathcal{T}$). Now each non-negative measurable function f is the limit of bounded non-negative measurable functions ($f = \lim_n f \wedge n$), and thus another application of the MCT shows that every $f \in m(\mathcal{S} \otimes \mathcal{T})^+$ satisfies (i) and (ii).

Now drop the assumption that μ is a finite measure. Because μ is σ -finite, we can choose $A_n \uparrow S$ such that $\mu A_n < \infty$. The measures $\mu_n = I_{A_n} \cdot \mu$ are finite measures, and thus each map $t \mapsto \int f(s, t) \mu_n(ds)$ is \mathcal{T} -measurable (where $f \geq 0$). Since $\int f(s, t) \mu(ds) = \lim_n \int f(s, t) \mu_n(ds)$, the MCT implies that the result holds for μ .

+

We now know that it is possible to define $\mu \otimes \nu$ in the ways indicated. What we don't (yet) know is that these constructions define a *measure*, and that they *coincide*.

For definiteness, we fix one of the above definitions:

Definition 16.3.8 Suppose that (S, \mathcal{S}, μ) and (T, \mathcal{T}, ν) are σ -finite measure spaces. Define a map $\mu \otimes \nu : \mathcal{S} \otimes \mathcal{T} \rightarrow \mathbb{R}^+$ by

$$(\mu \otimes \nu)B := \iint I_B(s, t) \nu(dt) \mu(ds) = \mu^s(\nu^t I_B(s, t)) \quad B \in \mathcal{S} \otimes \mathcal{T}$$

$\mu \otimes \nu$ is called the *product measure* of μ, ν .

□

Exercise 16.3.9 Show that $\mu \otimes \nu$ defines a σ -finite measure on $(S \times T, \mathcal{S} \otimes \mathcal{T})$.

□

The next two results show that (modulo certain conditions) we can calculate the integral w.r.t. $\mu \otimes \nu$ as a double integral, and the order of integration doesn't matter:

$$\int f d(\mu \otimes \nu) = \iint f(s, t) \nu(dt) \mu(ds) = \iint f(s, t) \mu(ds) \nu(dt)$$

We first show this for non-negative measurable functions:

Theorem 16.3.10 (Tonelli)

Suppose that (S, \mathcal{S}, μ) and (T, \mathcal{T}, ν) are σ -finite measure spaces. If $f \in m(\mathcal{S} \otimes \mathcal{T})^+$, then

$$(\mu \otimes \nu)f = \mu^s(\nu^t f(s, t)) = \nu^t(\mu^s f(s, t)) \quad (*)$$

Proof: We use the Monotone Class Theorem (Thm. 8.1.5). First assume that μ, ν are finite measures. The result is obvious if $f = I_{A \times B}$, where $A \times B$ measurable rectangle, (or cf. Exercise 16.3.6). The class

$$\mathcal{H} = \{f \in m(\mathcal{S} \otimes \mathcal{T}) : f \text{ is bounded and satisfies } (*)\}$$

is easily seen to satisfy the requirements of Thm. 8.1.3, and thus implies that \mathcal{H} contains every bounded $\mathcal{S} \otimes \mathcal{T}$ -measurable function. The result for arbitrary non-negative f follows by MCT.

A standard argument lifts the result to the case where μ, ν are merely σ -finite.

+

As a by-product, we obtain the result that our two possible definitions of $\mu \otimes \nu$ as iterated integrals coincide: If $B \in \mathcal{S} \otimes \mathcal{T}$, then I_B is a non-negative measurable function, and we may apply Tonelli's Thm.

For non-negative functions f , the integral μf always makes sense, but we may have $\mu f = \infty$. For arbitrary measurable f , we have to be more careful.

Theorem 16.3.11 (Fubini)

Suppose that (S, \mathcal{S}, μ) and (T, \mathcal{T}, ν) are σ -finite measure spaces. If $f \in \mathcal{L}^1(S \times T, \mathcal{S} \otimes \mathcal{T}, \mu \otimes \nu)$, then

$$(\mu \otimes \nu)f = \mu^s(\nu^t f(s, t)) = \nu^t(\mu^s f(s, t))$$

Here the map $t \mapsto \mu^s f(s, t)$ belongs to $\mathcal{L}^1(T, \mathcal{T}, \nu)$ for ν -a.e. $t \in T$. Similarly, the map $s \mapsto \nu^t f(s, t)$ belongs to $\mathcal{L}^1(S, \mathcal{S}, \mu)$ for μ -a.e. $s \in S$.

Proof: The result holds for $|f|$, by Tonelli's Thm., and hence $N_S = \{s \in S : \nu^t |f(s, t)| = +\infty\}$ is μ -null, and $N_T = \{t \in T : \mu^s |f(s, t)| = +\infty\}$ is ν -null. Redefine $f(s, t)$ to be zero when either $s \in N_S$ or $t \in N_T$; this won't affect the integral of f , by Thm. 6.3.9. The result follows by splitting f into positive and negative parts.

+

Remarks 16.3.12 (a) Fubini's Theorem allows the interchange of the order of integration, provided the integrand is integrable w.r.t the product measure. It follows from Fubini's Theorem that

$$\int \left(\int f \, d\nu \right) d\mu = \int \left(\int f \, d\mu \right) d\nu$$

provided that $f \in \mathcal{L}^1$. See Exercise 16.3.13 for what can happen if $f \notin \mathcal{L}^1$.

(b) Fubini's Theorem also easily extends to arbitrary finite products: If $(S_i, \mathcal{S}_i, \mu_i)$ are σ -finite measure spaces for $i = 1, \dots, n$, then

- (i) $\mathcal{S}_1 \otimes \dots \otimes \mathcal{S}_n$ is the σ -algebra on $S_1 \times \dots \times S_n$ which is generated by the projections $\pi_i : S_1 \times \dots \times S_n \rightarrow S_i : (s_1, \dots, s_n) \mapsto s_i$. It is also generated by the family of measurable "rectangles" $\mathcal{R} = \{A_1 \times \dots \times A_n : A_i \in \mathcal{S}_i \text{ for } i = 1, \dots, n\}$.
- (ii) $\mu_1 \otimes \dots \otimes \mu_n$ is the unique measure on $\mathcal{S}_1 \otimes \dots \otimes \mathcal{S}_n$ which assigns to every rectangle the measure

$$(\mu_1 \otimes \dots \otimes \mu_n)(A_1 \times \dots \times A_n) = \mu_1 A_1 \cdot \dots \cdot \mu_n A_n$$

- (iii) Fubini's Theorem states that if $f : S_1 \times \dots \times S_n \rightarrow \bar{\mathbb{R}}$ is $\mu_1 \otimes \dots \otimes \mu_n$ -integrable, then

$$\int_{S_1 \times \dots \times S_n} f \, d(\mu_1 \otimes \dots \otimes \mu_n) = \int_{S_1} \left(\int_{S_2} \dots \left(\int_{S_n} f \, d\mu_n \right) \dots d\mu_2 \right) d\mu_1$$

and that any interchange of the order of integration is permissible.

□

Exercise 16.3.13 Let

$$f(x, y) = \frac{x^2 - y^2}{(x^2 + y^2)^2}$$

Show that

$$\int_0^1 \int_0^1 f(x, y) \lambda(dy) \lambda(dx) = \frac{\pi}{4} \quad \int_0^1 \int_0^1 f(x, y) \lambda(dx) \lambda(dy) = -\frac{\pi}{4}$$

What can you conclude about

$$\int_{[0,1] \times [0,1]} f d(\lambda \otimes \lambda)$$

□

Appendix A

Convergence in \mathbb{R}

Some of the notions described in this chapter should be thoroughly familiar already. We include them here as preparation for the related, but more demanding, notions that we will encounter in the future.

A.1 Definition of Convergence

Let P be a property that a real number may (or may not) have. We write $P(x)$ if x has the property P . [For example, P could be the property of being positive, so that $P(1.23)$, but $\neg P(-\pi)$. Or Q could be the property of being irrational, in which case $\neg Q(1.23)$, whereas $Q(-\pi)$.] Now suppose that $\langle x_n \rangle_n$ is a sequence in \mathbb{R} , and that P is a property:

- We say $\langle x_n \rangle_n$ has property P *infinitely often* iff there are infinitely many n for such that $P(x_n)$ is true.
- We say $\langle x_n \rangle_n$ has property P *eventually* iff $P(x_n)$ is true for all n from some point onwards.

This can be made precise:

$$\begin{aligned}\langle x_n \rangle_n \text{ has property } P \text{ infinitely often} &\iff \forall n \exists m \geq n P(x_m) \\ \langle x_n \rangle_n \text{ has property } P \text{ eventually} &\iff \exists n \forall m \geq n P(x_m)\end{aligned}$$

Exercise A.1.1 Show that $\neg(\forall x P(x)) \iff \exists x(\neg P(x))$ and that $\neg(\exists x P(x)) \iff \forall x(\neg P(x))$. Conclude that $\neg(P(x_n), \text{i.o.}) \iff (\neg P(x_n), \text{ev.})$.

□

We first recall what it means for a sequence $\langle x_n \rangle_n$ of non-negative real numbers to converge to zero:

To say that $x_n \rightarrow 0$ means that $\langle x_n \rangle$ is “small” eventually.

The notion “small” is subjective, so we will demand that it holds for absolutely anybody’s idea of “small”. Specifically, suppose you define “small” by specifying some number $\varepsilon > 0$ and saying “A non-negative number x is small iff $x < \varepsilon$ ”. To say that $\langle x_n \rangle_n$ is eventually small then means that from some point onwards all the x_n ’s are small, i.e

$$\exists N \forall n \geq N [x_n < \varepsilon]$$

This must be true no matter what gauge $\varepsilon > 0$ of “smallness” you use. Thus:

<p>If $\langle x_n \rangle_n$ is a sequence of non-negative real numbers, we say</p> $x_n \rightarrow 0 \quad \Longleftrightarrow \quad \forall \varepsilon > 0 \exists N \forall n \geq N [x_n < \varepsilon]$
--

Thus $x_n \rightarrow 0$ iff given any $\varepsilon > 0$ it is possible to find a natural number N such that

$$x_n < \varepsilon \quad \text{whenever } n \geq N$$

The number N typically depends on ε . The smaller $\varepsilon > 0$, the greater N usually has to be.

It is now simple to define convergence of arbitrary sequences in \mathbb{R} : To say that $x_n \rightarrow x$ means that the distance between x_n and x converges to 0, i.e.

$$x_n \rightarrow x \quad \Leftrightarrow \quad |x_n - x| \rightarrow 0$$

Now the distance $|x_n - x|$ between x_n and x is non-negative, so we already know what $|x_n - x| \rightarrow 0$ means: It means $\forall \varepsilon > 0 \exists N \forall n \geq N [|x_n - x| < \varepsilon]$. Thus

<p>If $\langle x_n \rangle_n$ is a sequence of real numbers, we say</p> $x_n \rightarrow x \quad \Longleftrightarrow \quad \forall \varepsilon > 0 \exists N \forall n \geq N [x_n - x < \varepsilon]$

We also write

$$x = \lim_n x_n \quad \text{for} \quad x_n \rightarrow x$$

Thus $x_n \rightarrow x$ iff given any $\varepsilon > 0$ it is possible to find a natural number N such that

$$|x_n - x| < \varepsilon \quad \text{whenever } n \geq N$$

The number N typically depends on ε . The smaller $\varepsilon > 0$, the greater N usually has to be.

We also sometimes say that a sequence converges to $\pm\infty$.

<p>To say that $x_n \rightarrow \infty$ means $\langle x_n \rangle_n$ is “large” eventually.</p>
--

The notion “large” is subjective, so we will demand that it holds for absolutely anybody’s idea of “large”. Specifically, suppose you define “large” by specifying some number $K > 0$ and saying “A number x is large iff $x > K$ ”. To say that $\langle x_n \rangle_n$ is eventually large then means that from some point onwards all the x_n ’s are large, i.e

$$\exists N \forall n \geq N [x_n > K]$$

This must be true no matter what gauge $K > 0$ of “largeness” you use. Thus:

If $\langle x_n \rangle_n$ is a sequence of real numbers, we say

$$x_n \rightarrow \infty \iff \forall K > 0 \exists N \forall n \geq N [x_n > K]$$

We say that $x_n \rightarrow -\infty$ iff $-x_n \rightarrow +\infty$.

When $x_n \rightarrow \pm\infty$, then $\lim_n x_n$ does *not* exist (as a real number). We then say that $\lim_n x_n$ exists in the *extended sense*.

The next theorem is left as an exercise:

Theorem A.1.2 (Sandwich Theorem)

Suppose that $\langle x_n \rangle_n, \langle y_n \rangle_n$ and $\langle z_n \rangle_n$ are sequences in \mathbb{R} which satisfy the following conditions:

(i) $x_n \leq y_n \leq z_n$ for all $n \in \mathbb{N}$ (or merely eventually);

(ii) There is $l \in \mathbb{R}$ such that $x_n \rightarrow l$ and $z_n \rightarrow l$.

Then also $y_n \rightarrow l$.

Exercise A.1.3 The aim of this exercise is to prove the Sandwich Theorem. So let $\varepsilon > 0$. We must show that there is $N \in \mathbb{N}$ such that $|y_n - l| < \varepsilon$ whenever $n > N$, or equivalently, that $l - \varepsilon < y_n < l + \varepsilon$ whenever $n > N$.

- Assume first that $x_n \leq y_n \leq z_n$ for all $n \in \mathbb{N}$. Explain why there is $N_1 \in \mathbb{N}$ such that whenever $n > N_1$, we have $l - \varepsilon < x_n < l + \varepsilon$.
- Now explain why there is an $N \in \mathbb{N}$ such that whenever $n > N$, we have *both* $l - \varepsilon < x_n < l + \varepsilon$ and $l - \varepsilon < z_n < l + \varepsilon$.
- Now explain why also $l - \varepsilon < y_n < l + \varepsilon$ whenever $n > N$.
- The Theorem has now been proved for the case where $x_n \leq y_n \leq z_n$ for all $n \in \mathbb{N}$. Modify your proof slightly to show that the Theorem remains true if we have $x_n \leq y_n \leq z_n$ *eventually*.

□

You should also be familiar with the following theorem, whose proof we leave as an exercise. (It can be found in any introductory text on Real Analysis.)

Theorem A.1.4 Let $\langle x_n \rangle_n, \langle y_n \rangle_n$ be sequences in \mathbb{R} , with $x_n \rightarrow x$, $y_n \rightarrow y$. Also let $\alpha \in \mathbb{R}$. Then

- (a) $(x_n + y_n) \rightarrow x + y$;
- (b) $\alpha x_n \rightarrow \alpha x$;
- (c) $x_n y_n \rightarrow xy$;
- (d) $\frac{1}{x_n} \rightarrow \frac{1}{x}$ (provided $x_n \neq 0$ for $n \in \mathbb{N}$, and $x \neq 0$);
- (e) $\frac{x_n}{y_n} \rightarrow \frac{x}{y}$ (provided $y_n \neq 0$ for $n \in \mathbb{N}$, and $y \neq 0$).

A.2 The Completeness Axiom

Much of the power of real analysis comes from the ability to construct objects as certain limits. That these limits exist is a consequence of the *completeness axiom*. In order to state this axiom, we need some preliminary definitions:

Definition A.2.1 Let $A \subseteq \mathbb{R}$.

- (a) We say that $u \in \mathbb{R}$ is an *upper bound* for A iff every element of A is $\leq u$ (i.e. iff $\forall a \in A [a \leq u]$).
- (b) We say that $A \subseteq \mathbb{R}$ is *bounded above* iff A has a (finite) upper bound.
- (c) Similarly, we say that $l \in \mathbb{R}$ is a *lower bound* of $A \subseteq \mathbb{R}$ iff $\forall a \in A (l \leq a)$. If A has a (finite) lower bound, it is *bounded below*. A set which is both bounded above and below is said to be *bounded*.
- (d) We say that u_0 is the *supremum* (or *least upper bound*) of $A \subseteq \mathbb{R}$ iff
 - (i) u_0 is an upper bound of A , and
 - (ii) If u is an upper bound of A , then $u_0 \leq u$ (i.e. u_0 is an upper bound which is \leq any other upper bound).

We write this as

$$u_0 = \sup A$$

- (e) Similarly, we say that l_0 is the *infimum* (or *greatest bound*) of $A \subseteq \mathbb{R}$ iff
 - (i) l_0 is a lower bound of A , and
 - (ii) If l is an upper bound of A , then $l_0 \geq l$ (i.e. l_0 is a lower bound which is \geq any other lower bound).

We write this as

$$l_0 = \inf A$$

Remarks A.2.2 1. Upper bounds, if they exist, are not unique: If u is an upper bound of A and $v \geq u$, then v is also an upper bound of A .

2. The notions of \sup , \inf generalize the notions of \max , \min : If $x = \max A$, then $x = \sup A$. However, a set $A \subseteq \mathbb{R}$ may have a supremum without having a maximum.
3. The following statements are obvious, but often useful:

$$x < \sup A \quad \Leftrightarrow \quad \exists a \in A [x < a]$$

$$x > \inf A \quad \Leftrightarrow \quad \exists a \in A [x > a]$$

4. If A has no (finite) upper bound, we may write $\sup A = \infty$.
5. If $A \subseteq \mathbb{R}$, then $\sup(-A) = -\inf A$ (where $-A := \{-a : a \in A\}$).

□

Here is the fundamental axiom of analysis:

Completeness Axiom *Every non-empty subset of \mathbb{R} which has an upper bound has a least upper bound, i.e. if $A \subseteq \mathbb{R}$ is bounded above then $\sup A$ exists.*

Using Remarks A.2.2.5, it is easy to see that every subset of \mathbb{R} which has a lower bound has a greatest lower bound.

Remarks A.2.3 1. Note that the completeness axiom fails for \mathbb{Q} : It is not true that every bounded set of rational numbers has a *rational* least upper bound. To see this, take

$$A := \{x \in \mathbb{Q} : x^2 < 2\}$$

This set does have supremum in \mathbb{R} , namely $\sup A = \sqrt{2}$. However, $\sqrt{2}$ is irrational.

2. Why should we believe this axiom to be true for \mathbb{R} ? Here is a thought experiment that provides some intuition: Suppose that A is a set of *non-negative reals*. For each $a \in A$, take a line segment of length a . Stack all these line segments on top of one another, on a blank page, with bottom points aligned. Since the line segments have zero thickness, all you will see is a single line segment. This “line segment” ought to have a length, and a little thought will convince you that this length is $\sup A$.
3. If A is non-empty, we obviously have $\inf A \leq \sup A$. Yet

$$\inf \emptyset = \infty \quad \sup \emptyset = -\infty$$

(Explain!)

□

The following theorem is basic: Recall that a sequence $\langle x_n \rangle_n$ of real numbers is increasing if and only if $x_1 \leq x_2 \leq x_3 \leq \dots$, i.e. iff $m \leq n \Rightarrow x_m \leq x_n$.

Theorem A.2.4 If $\langle x_n \rangle$ is an increasing sequence of real numbers which is bounded above, then $\langle x_n \rangle_n$ converges, and

$$\lim_n x_n = \sup_n x_n$$

Proof: Because $\{x_n : n \in \mathbb{N}\}$ is bounded above $x := \sup\{x_n : n \in \mathbb{N}\}$ exists. We show $x_n \rightarrow x$.

Let $\varepsilon > 0$. Then $x - \varepsilon < \sup\{x_n : n \in \mathbb{N}\}$. Hence there is $N \in \mathbb{N}$ such that $x - \varepsilon < x_N$ (cf. Remarks A.2.2.3). If $n \geq N$ then $x_n \geq x_N$, and hence $x - \varepsilon < x_n \leq x$. (All the x_n are $\leq x$, because x is an upper bound of the x_n .) It follows that $|x_n - x| = x - x_n < \varepsilon$ whenever $n \geq N$.

□

By Remarks A.2.2.5 it follows that every decreasing sequence $\langle x_n \rangle_n$ which is bounded below has a limit, and that this limit is $\inf_n x_n$.

Remarks A.2.5 If $\langle x_n \rangle_n$ is an increasing sequence which converges to x , we often write

$$x_n \uparrow x \quad \text{or} \quad x = \uparrow \lim_n x_n$$

instead of $x_n \rightarrow x$. Similarly, if $\langle x_n \rangle_n$ is a decreasing convergent sequence, we write

$$x_n \downarrow x \quad \text{or} \quad x = \downarrow \lim_n x_n$$

□

Note that the preceding theorem guarantees that a bounded monotone sequence has a limit, even if we have no way of directly assessing what that limit might actually be. This is the case in the following example:

Example A.2.6 Define $x_n = (1 + \frac{1}{n})^n$. We show that $\langle x_n \rangle_n$ converges. By the preceding theorem, it suffices to show (i) that $\langle x_n \rangle_n$ is increasing, and (ii) that it is bounded. Now by the Binomial Theorem

$$\begin{aligned} x_n &= 1 + \frac{n}{1} \frac{1}{n} + \frac{n(n-1)}{2!} \frac{1}{n^2} + \cdots + \frac{n(n-1)\cdots 2 \cdot 1}{n!} \frac{1}{n^n} \\ &= 1 + 1 + \frac{1}{2!} \left(1 - \frac{1}{n}\right) + \cdots + \frac{1}{n!} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{n-1}{n}\right) \end{aligned}$$

Similarly,

$$x_{n+1} = 1 + 1 + \frac{1}{2!} \left(1 - \frac{1}{n+1}\right) + \cdots + \frac{1}{(n+1)!} \left(1 - \frac{1}{n+1}\right) \left(1 - \frac{2}{n+1}\right) \cdots \left(1 - \frac{n}{n+1}\right)$$

Now we compare terms. x_n has $n+1$ terms, whereas, x_{n+1} has $n+2$ terms, all non-negative. x_n and x_{n+1} agree on the first two terms. Now if $2 < k \leq n+1$, then the k^{th} term of x_n is

$$\frac{1}{(k-1)!} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{k-2}{n}\right)$$

whereas the k^{th} term of x_{n+1} is

$$\frac{1}{(k-1)!} \left(1 - \frac{1}{n+1}\right) \left(1 - \frac{2}{n+1}\right) \cdots \left(1 - \frac{k-2}{n+1}\right)$$

It is therefore clear that the k^{th} term of x_n is less than the k^{th} term of x_{n+1} . Moreover, x_{n+1} has one more term, which is strictly positive. It follows that $x_n < x_{n+1}$.

Thus $\langle x_n \rangle_n$ is an increasing sequence.

Next, we show that $\langle x_n \rangle_n$ is bounded. Look again at the k^{th} term of x_n : We have

$$\frac{1}{(k-1)!} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{k-2}{n}\right) \leq \frac{1}{2^{k-2}}$$

This is (i), because

$$2^{k-2} = 2 \cdot 2 \cdots 2 \leq 2 \cdot 3 \cdots (k-1)$$

and (ii), because

$$\left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{k-2}{n}\right) \leq 1 \cdot 1 \cdots 1$$

It follows that

$$x_n \leq 1 + 1 + \frac{1}{2} + \frac{1}{2^2} + \cdots + \frac{1}{2^{n-1}} \leq 3$$

and thus that $x_n \leq 3$ for all n . Hence $\langle x_n \rangle_n$ is bounded.

We can now conclude that $\langle x_n \rangle_n$ converges, though we do not yet know precisely where it converges to.

If you stick the x_n into a calculator, you will see that $x_n \rightarrow e$, where $e = 2.7182818\ldots$ is the base of the natural logarithm.

□

A.3 \limsup and \liminf ; Subsequences

Suppose that $\langle x_n \rangle_n$ is a *bounded* sequence in \mathbb{R} . Construct two new sequences as follows:

$$y_n = \sup\{x_m : m \geq n\} \quad z_n = \inf\{x_m : m \geq n\}$$

Because $\langle x_n \rangle$ is bounded, y_n and z_n exist (i.e. are finite real numbers), by the Completeness Axiom.

Suppose, for example, that $x_n = \frac{(-1)^n}{n}$ for $n \geq 1$. Then

$$y_1 = \sup\{-1, \frac{1}{2}, -\frac{1}{3}, \frac{1}{4}, -\frac{1}{5}, \dots\} = \frac{1}{2}$$

$$y_2 = \sup\{\frac{1}{2}, -\frac{1}{3}, \frac{1}{4}, -\frac{1}{5}, \dots\} = \frac{1}{2}$$

$$y_3 = \sup\{-\frac{1}{3}, \frac{1}{4}, -\frac{1}{5}, \dots\} = \frac{1}{4}$$

$$y_4 = \sup\{\frac{1}{4}, -\frac{1}{5}, \dots\} = \frac{1}{4}$$

i.e. $\langle y_n \rangle$ is the sequence $\frac{1}{2}, \frac{1}{2}, \frac{1}{4}, \frac{1}{4}, \frac{1}{6}, \frac{1}{6}, \frac{1}{8}, \dots$

Similarly, you can check that $\langle z_n \rangle$ is the sequence $-1, \frac{1}{3}, -\frac{1}{3}, -\frac{1}{5}, -\frac{1}{5}, -\frac{1}{7}, -\frac{1}{7}, -\frac{1}{9}, \dots$

Exercise A.3.1 (a) Given $\langle x_n \rangle_n$, write down the first 6 terms of $y_n = \sup\{x_m : m \geq n\}$ and $z_n = \inf\{x_m : m \geq n\}$.

(i) $x_n = (-1)^n$

(ii) $x_n = \frac{1}{n}$

(iii) $x_n = \begin{cases} 1 + \frac{1}{n} & \text{if } n \text{ is odd} \\ -1 - 2^{-n} & \text{if } n \text{ is even} \end{cases}$

(iv) $x_n = \begin{cases} 1 - \frac{1}{n} & \text{if } n \text{ is odd} \\ -1 + 2^{-n} & \text{if } n \text{ is even} \end{cases}$

(b) Note that each of the above sequences $\langle y_n \rangle$ is decreasing, and that each of the $\langle z_n \rangle_n$ is increasing. Can you explain why?

(c) Finally, since the $\langle y_n \rangle$ and $\langle z_n \rangle$ are bounded monotone sequences, they must converge (by Theorem A.2.4). Write down $\lim_n y_n$ and $\lim_n z_n$ for each of the sequences in (a)(i)-(iv).

□

As noted in the above exercise, $\langle y_n \rangle$ is a decreasing sequence, and $\langle z_n \rangle$ is increasing. To see this, let $A_n = \{x_m : m \geq n\}$. Clearly

$$A_1 \supseteq A_2 \supseteq A_3 \supseteq \dots$$

Hence

$$\sup A_1 \geq \sup A_2 \geq \sup A_3 \geq \dots \quad \text{and} \quad \inf A_1 \leq \inf A_2 \leq \inf A_3 \leq \dots$$

(Note that if $A \subseteq B$, then $\sup A \leq \sup B$, and $\inf A \geq \inf B$.)

Since $y_n = \sup A_n$ and $z_n = \inf A_n$, we see that

$$y_1 \geq y_2 \geq y_3 \geq \dots \quad \text{and} \quad z_1 \leq z_2 \leq z_3 \leq \dots$$

Now any bounded monotone sequence converges (Theorem A.2.4), and thus $\lim_n y_n$ and $\lim_n z_n$ exist if $\langle x_n \rangle$ is bounded. We now define $\limsup_n x_n = \lim_n y_n$, and $\liminf_n x_n = \lim_n z_n$:

Definition A.3.2 Let $\langle x_n \rangle$ be a sequence in \mathbb{R} . We define the *limit superior* of $\langle x_n \rangle$ by

$$\limsup_n x_n = \lim_{n \rightarrow \infty} \sup_{m \geq n} x_m$$

where we adopt the convention that if $\langle x_n \rangle$ is not bounded above, we set $\limsup_n x_n = +\infty$. Similarly, we define the *limit inferior* of $\langle x_n \rangle$ by

$$\liminf_n x_n = \lim_{n \rightarrow \infty} \inf_{m \geq n} x_m$$

where we adopt the convention that if $\langle x_n \rangle$ is not bounded below, we set $\liminf_n x_n = -\infty$.

Because $\langle y_n \rangle$ is decreasing, we have $\limsup_n x_n = \downarrow \lim_n y_n = \inf_n y_n$ (by Theorem A.2.4). Hence (and similarly)

$$\limsup_n x_n = \inf_n \sup_{m \geq n} x_m \quad \liminf_n x_n = \sup_n \inf_{m \geq n} x_m$$

The notions of lim sup and lim inf are often regarded as quite difficult, so we will try to improve our understanding of it. Note that since

$$\liminf_n x_n = \sup_n \inf_{m \geq n} x_m = -\inf_n \sup_{m \geq n} (-x_m) = -\limsup_n (-x_n)$$

(because $-\sup A = \inf(-A)$), we need to prove a result only for lim sup in order to obtain immediately a corresponding result for lim inf.

Proposition A.3.3 *If $\langle x_n \rangle$ is bounded above, then*

(a)

$$x < \limsup_n x_n \implies x < x_n \text{ infinitely often}$$

(b)

$$x \leq x_n \text{ infinitely often} \implies x \leq \limsup_n x_n$$

(c)

$$x > \limsup_n x_n \implies x > x_n \text{ eventually}$$

(d)

$$x \geq x_n \text{ eventually} \implies x \geq \limsup_n x_n$$

(e)

$$\limsup_n x_n = \sup\{x : x_n > x \text{ infinitely often}\} = \inf\{x : x_n \leq x \text{ eventually}\}$$

The proof is an exercise:

Exercise A.3.4 We prove Propn. A.3.3.

Define $y_n := \sup\{x_m : m \geq n\}$, and let $\bar{x} := \limsup_n x_n = \downarrow \lim_n y_n$.

(a) Suppose that $x < \bar{x}$. Show that $y_n > x$ for all n . Conclude that $\forall n \exists m \geq n (x_m > x)$.

(b) Suppose that $x \leq x_n$ i.o.. Explain why $y_n \geq x$ for all n . Conclude that $\bar{x} \geq x$.

(c) Use (b) and some logic: $x > \bar{x}$ implies $x \not\leq \bar{x}$, so $x \not\leq x_n$ i.o. and hence $x > x_n$ ev..

(d) Use (a) and logic.

(e) Let $A := \{x : x_n > x \text{ i.o.}\}$. Use (b) to show $\sup A \leq \bar{x}$. Use (a) to show $\sup A \not\leq \bar{x}$ by considering an x such that $\sup A < x < \bar{x}$.

Proposition A.3.5 *Suppose that $\langle x_n \rangle_n, \langle y_n \rangle_n$ are bounded sequences in \mathbb{R} , and that $\lambda \in \mathbb{R}$.*

(a) $\liminf_n x_n \leq \limsup_n x_n$

(b) *If $\lambda \geq 0$, then $\limsup_n \lambda x_n = \lambda \limsup_n x_n$, and $\liminf_n \lambda x_n = \lambda \liminf_n x_n$*

(c) *If $\lambda < 0$, then $\limsup_n \lambda x_n = \lambda \liminf_n x_n$, and $\liminf_n \lambda x_n = \lambda \limsup_n x_n$*

(d) $\limsup_n (x_n + y_n) \leq \limsup_n x_n + \limsup_n y_n$

(e) $\liminf_n (x_n + y_n) \geq \liminf_n x_n + \liminf_n y_n$

(f) *If $x_n \leq y_n$, then $\limsup_n x_n \leq \limsup_n y_n$ and $\liminf_n x_n \leq \liminf_n y_n$*

Proof: Here's a proof of (a): Suppose that $z < \liminf_n x_n$ then $x_n > z$ eventually, so also $x_n > z$ infinitely often. It follows that $z \leq \limsup_n x_n$. Hence¹ $\limsup_n x_n \geq z$ whenever $z < \liminf_n x_n$, and thus $\limsup_n x_n \geq \liminf_n x_n$ as well.

The rest of this proposition is left as an exercise.

¹ What we are using here is that, if $a \geq x$ whenever $x < b$, then also $a \geq b$. For suppose that $a < b$. Choose x such that $a < x < b$. Then $x < b$, so $a \geq x$ — contradiction, since $a < x$. Hence $a \not< b$, i.e. $a \geq b$.

⊥

Exercise A.3.6 Prove the remainder of Proposition A.3.5.

[Hints: (c) If $x_n > z$ infinitely often and $\lambda < 0$, then $\lambda x_n < \lambda z$ infinitely often.

(d) If $z > \limsup_n x_n$ and $w > \limsup_n y_n$, then $x_n < z$ eventually and $y_n < w$ eventually. Hence $x_n + y_n < z + w$ eventually.]

□

We are still not done with lim sup and lim inf. It is also possible to find a characterization in terms of *subsequences*. Roughly speaking, if you write down all the terms of a sequence $\langle x_n \rangle_n$, and then delete some of these terms, what remains is a subsequence. (However, you can't delete so many terms that there are only finitely many left.)

This is best understood by looking at some examples: The sequence $2, 3, 5, 7, 11, \dots$ of primes is a subsequence of the sequence $1, 2, 3, 4, \dots$ of natural numbers:

$$\cancel{1}, 2, 3, \cancel{4}, 5, \cancel{6}, 7, \cancel{8}, \emptyset, \cancel{10}, 11, \dots$$

In the subsequence, the order of elements remains the same as what it was in the original: 2 comes before 3 comes before 5... etc. in both sequences.

The sequence $3, 2, 6, 5, 9, 8, \dots$ is *not* a subsequence of $1, 2, 3, 4, 5, \dots$. Not only have we deleted all numbers of the form $3n - 2$, we have also rearranged them so that $3n$ is before $3n - 1$. In the sequence of natural numbers, 2 is before 3, but in this new sequence, 3 is before 2. Such rearrangements are not allowed when you construct a subsequence.

The following definition should now make sense:

Definition A.3.7 Let $\langle x_n \rangle_n$ be a sequence in \mathbb{R} , and suppose that $\langle n_k \rangle_k$ is a strictly increasing sequence in \mathbb{N} (i.e. $n_1 < n_2 < n_3 < \dots$). Then the sequence

$$\langle x_{n_k} \rangle_k = x_{n_1}, x_{n_2}, x_{n_3}, \dots$$

is called a *subsequence* of $\langle x_n \rangle_n$.

For example

$$\begin{aligned}\langle x_{2n} \rangle_n &= x_2, x_4, x_6, \dots \\ \langle x_{3n-1} \rangle_n &= x_2, x_5, x_8, \dots \\ \langle x_{5^n} \rangle_n &= x_5, x_{25}, x_{125}, \dots\end{aligned}$$

are subsequences of $\langle x_n \rangle_n$.

Remarks A.3.8 1. One easy but useful fact to note is the following: If $n_1 < n_2 < n_3 < \dots$ is a strictly increasing sequence of natural numbers, then $n_k \geq k$ (for each $k \in \mathbb{N}$).

If you can't see this immediately, try proving it by induction. Clearly $n_1 \geq 1$. Now suppose that $n_k \geq k$. Then $n_{k+1} > n_k \geq k$, and thus $n_{k+1} \geq k + 1$.

2. Note that the n in $\langle x_n \rangle_n$ is a “dummy” variable — not really a variable at all. This means that it doesn't matter if we replace the n by some other symbol k : $\langle x_k \rangle_k$ is *exactly the same* as $\langle x_n \rangle_n$.

For example $\langle \frac{1}{k} \rangle_k = 1, \frac{1}{2}, \frac{1}{3}, \dots = \langle \frac{1}{n} \rangle_n$.

In particular, $\lim_k x_k$ is exactly the same as $\lim_n x_n$, $\sup_k x_k$ the same as $\sup_n x_n$, etc.

In the expression $\langle x_n \rangle_n$, the variable n is a *bound* variable, constrained to take on *all* possible values in the set \mathbb{N} . We have a similar situation when we deal with definite integrals: The expression $\int_0^1 x \, dx$ is a number, namely $\frac{1}{2}$, and not a variable, even though it seems to have a variable x occurring in it. However, *that* x is a bound variable, constrained to take on all possible values between 0 and 1. It doesn't matter if we replace *the* x by some other symbol u : $\int_0^1 x \, dx$ is *exactly the same* as $\int_0^1 u \, du$

□

One important type of subsequence is a *tail sequence*. A tail sequence of $\langle x_n \rangle_n$ is a subsequence which consists of all terms of x_n from some N onwards, e.g. $5, 6, 7, \dots$ is a tail sequence of $1, 2, 3, \dots$. Similarly $\frac{1}{100}, \frac{1}{101}, \frac{1}{102}, \dots$ is a tail sequence of $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \dots$. Thus $\langle y_n \rangle_n$ is a tail sequence of $\langle x_n \rangle_n$ iff there is an integer $N \geq 0$ such that $y_n = x_{N+n}$.

Example A.3.9 If

$$x_n = \begin{cases} \frac{1}{n} & \text{if } n \text{ is odd} \\ 1 + \frac{1}{n^2} & \text{if } n \text{ is even} \end{cases}$$

then $\langle x_n \rangle_n$ is divergent. However, the sequences $\langle y_n \rangle_n, \langle z_n \rangle_n$ defined by

$$y_n = x_{2n-1}, \quad z_n = x_{2n}$$

are convergent subsequences of $\langle x_n \rangle_n$, with $y_n \rightarrow 0$, and $z_n \rightarrow 1$.

If you think long enough, it should be clear that a subsequence $\langle x_{n_k} \rangle_k$ of $\langle x_n \rangle_n$ converges if and only if: EITHER the sequence $\langle n_k \rangle_k$ is odd eventually (in which case $x_{n_k} \rightarrow 0$), OR $\langle n_k \rangle_k$ is even eventually (in which case $x_{n_k} \rightarrow 1$).

Similarly, if $\langle n_k \rangle_k$ is BOTH odd infinitely often and even infinitely often, then $\langle x_{n_k} \rangle_k$ diverges.

□

The following proposition claims two things:

- If $\langle x_n \rangle_n$ is a convergent sequence, then every subsequence of $\langle x_n \rangle_n$ converges, and to the same limit.
- If a tail sequence of $\langle x_n \rangle_n$ is convergent, then $\langle x_n \rangle_n$ is itself convergent, and to the same limit.

Proposition A.3.10 (a) If $x_n \rightarrow x$, and if $\langle y_n \rangle_n$ is a subsequence of $\langle x_n \rangle_n$, then $y_n \rightarrow x$ as well.

(b) If $\langle y_n \rangle_n$ is a tail sequence of $\langle x_n \rangle_n$, and if $y_n \rightarrow x$, then also $x_n \rightarrow x$.

Exercise A.3.11 We prove Propn. A.3.10:

- (a) Suppose that $y_k = x_{n_k}$, where $n_1 < n_2 < n_3 < \dots$. We must show that $y_k \rightarrow x$, i.e. that for every $\varepsilon > 0$, there is a $K \in \mathbb{N}$ such that $|y_k - x| < \varepsilon$ whenever $k > K$.

So let $\varepsilon > 0$ be given. First explain why we can choose $M \in \mathbb{N}$ such that $|x_m - x| < \varepsilon$ whenever $m > M$. Now use Remarks A.3.8 to explain why we may choose $K \in \mathbb{N}$ such that $n_k > M$ whenever $k > K$. Finally, show that if $k > K$, we have

$$|y_k - x| = |x_{n_k} - x| < \varepsilon$$

and conclude.

- (b) Suppose that $\langle y_n \rangle_n$ is a convergent tail subsequence of $\langle x_n \rangle_n$, and that $y_n \rightarrow x$. We must show that also $x_n \rightarrow x$. So let $\varepsilon > 0$. Explain why there is N such that $n > N$ implies $|y_n - x| < \varepsilon$. Next explain why there is a non-negative integer M is such that $y_n = x_{n+M}$. Conclude that

$$|x_n - x| < \varepsilon \quad \text{whenever} \quad n > N + M$$

□

Example A.3.12 Consider again the sequence

$$x_n = \begin{cases} \frac{1}{n} & \text{if } n \text{ is odd} \\ 1 + \frac{1}{n^2} & \text{if } n \text{ is even} \end{cases}$$

The sequences

$$y_n = \frac{1}{2n+1} \quad z_n = 1 + \frac{1}{4n^2}$$

are subsequences of $\langle x_n \rangle_n$. Since $y_n \rightarrow 0$ and $z_n \rightarrow 1$, we can conclude that $\langle x_n \rangle_n$ is divergent. For if $\langle x_n \rangle_n$ converges (to x , say), then all its subsequences would also converge to the same limit x . But here we have two subsequences which converge to different limits.

□

The next exercise is preparation for Propn. A.3.14:

- Exercise A.3.13** (a) Let $x_n = (-1)^n + \frac{1}{n}$. Find a decreasing subsequence which converges to $\limsup_n x_n$. Is there an increasing subsequence which converges to $\limsup_n x_n$?
- (b) Let $x_n = (-1)^n - \frac{1}{n}$. Find an increasing subsequence which converges to $\limsup_n x_n$. Is there a decreasing sequence which converges to $\limsup_n x_n$?

□

Proposition A.3.14 Every sequence has a monotone subsequence.

In fact every sequence has a monotone subsequence which converges to $\limsup_n x_n$ (and similarly, one which converges to $\liminf_n x_n$).

Proof: Given a sequence $\langle x_n \rangle_n$, we show that there is a monotone subsequence which converges to $\limsup_n x_n$. Let $\bar{x} = \limsup_n x_n$, and put $y_n = \sup\{x_m : m \geq n\}$ (so that $y_n \downarrow \bar{x}$). We distinguish two cases.

Case 1: $\bar{x} < y_n$ for all n .

(We allow here the case $\bar{x} = -\infty$.) In that case, we can choose a decreasing subsequence $\langle x_{n_k} \rangle_k$ inductively, as follows: Let $N_1 = 1$. Since $y_{N_1} > \bar{x}$, there is $n_1 \geq N_1$ so that $x_{n_1} > \bar{x}$. Next, since $y_n \downarrow \bar{x}$, there is $N_2 > n_1$ such that $y_{N_2} < x_{n_1}$ (i.e. $x_m < x_{n_1}$ for all $m \geq N_2$). Since, by hypothesis $y_{N_2} > \bar{x}$, there is $n_2 \geq N_2$ such that $x_{n_2} > \bar{x}$ also. Thus $\bar{x} < x_{n_2} < x_{n_1}$.

Keep going in the same way: Once we have constructed $N_1 < N_2 < \cdots < N_k$ and $n_1 < n_2 < \cdots < n_k$ such that

$$N_{j+1} > n_j \geq N_j \quad \bar{x} < x_{n_{j+1}} < x_{n_j} \quad \text{for } j = 1, \dots, k-1$$

we may choose $N_{k+1} > n_k$ so that $y_{N_{k+1}} < x_{n_k}$. Since also $y_{N_{k+1}} > \bar{x}$, there is $n_{k+1} \geq N_{k+1}$ such that $x_{n_{k+1}} > \bar{x}$. Since $n_{k+1} \geq N_{k+1}$, we see that $n_{k+1} > n_k$. Since $x_{n_{k+1}} \leq y_{N_{k+1}}$, we have $\bar{x} < x_{n_{k+1}} < x_{n_k}$.

This completes the construction of the subsequence $\langle x_{n_k} \rangle_k$. Note that $\langle y_{N_k} \rangle_k$ is a subsequence of $\langle y_n \rangle_n$, so that $y_{N_k} \downarrow \bar{x}$. Since clearly $\bar{x} < x_{n_k} \leq y_{N_k}$ for all k , the Sandwich Theorem ensures that $x_{n_k} \downarrow \bar{x}$ as well.

Case 2: There is N_0 such that $y_{N_0} = \bar{x}$.

(We allow here the case $\bar{x} = +\infty$.) In that case, since y_n is a decreasing sequence converging to \bar{x} , we must have $y_n = \bar{x}$ for all $n \geq N_0$ also. In particular, it follows that $x_n \leq \bar{x}$ for all $n \geq N_0$. Thus either (i) $x_n < \bar{x}$ eventually, or (ii) $x_n = \bar{x}$ infinitely often. If (ii) holds, there is obviously a constant (hence monotone) subsequence converging to \bar{x} , so it remains to deal with (i).

Suppose therefore that $x_n < \bar{x}$ for all $n \geq N_1$, and let $N = \max\{N_0, N_1\}$. Then

$$\forall n \geq N \ (y_n = \bar{x} \wedge x_n < \bar{x})$$

Define $t_n = \bar{x} - \frac{1}{n}$ if \bar{x} is finite, and put $t_n = n$ if \bar{x} is infinite. The point is that $t_n \uparrow \bar{x}$, whether \bar{x} is finite or not. Inductively construct an increasing subsequence x_{n_k} as follows: Choose $n_1 = N$, so that $x_{n_1} < \bar{x}$. Because $y_n = \bar{x}$ for all $n \geq N$, there is $n_2 > n_1$ such that $\max\{x_{n_1}, t_2\} < x_{n_2}$ of course, also $x_{n_2} < \bar{x}$. Proceed in the same way: Given $n_1 < n_2 < \dots < n_k$ such that

$$\max\{x_{n_j}, t_{j+1}\} < x_{n_{j+1}} < \bar{x} \quad \text{for } j = 1, \dots, k-1$$

choose $n_{k+1} > n_k$ such that $x_{n_{k+1}} > \max\{x_{n_k}, t_{k+1}\}$ — this is possible because $y_{n_{k+1}} = \sup\{x_n : n > n_k\} = \bar{x}$.

In this way we obtain a strictly increasing subsequence $\langle x_{n_k} \rangle_k$ such that $t_k < x_{n_k} < \bar{x}$. Since $t_k \rightarrow \bar{x}$, we see that $x_{n_k} \rightarrow \bar{x}$ also, by the Sandwich theorem.

—

Here is our final characterization of \limsup and \liminf .

Proposition A.3.15 *Suppose that $\langle x_n \rangle$ is a bounded sequence. Then $\limsup_n x_n$ is the biggest number to which $\langle x_n \rangle$ has a convergent subsequence, and $\liminf_n x_n$ is the smallest number to which $\langle x_n \rangle_n$ has a convergent subsequence.*

Proof: By Proposition A.3.14, $\langle x_n \rangle_n$ does have a subsequence converging to $\limsup_n x_n$.

Now suppose that $\langle x_{n_k} \rangle_k$ is a subsequence of $\langle x_n \rangle$, and that $x_{n_k} \rightarrow a$, and let $\bar{x} = \limsup_n x_n$. If $a > \bar{x}$, we may choose $\varepsilon > 0$ such that $a - \varepsilon > \bar{x}$ also (e.g. $\varepsilon = \frac{1}{2}(a + \bar{x})$ will do). Since $x_{n_k} \rightarrow a$, eventually $x_{n_k} > a - \varepsilon$, i.e. there is K such that $x_{n_k} > a - \varepsilon$ whenever $k \geq K$. The set $\{x_{n_k} : k \geq K\}$ is therefore an infinite set of x_n 's which are greater than $a - \varepsilon$, and thus $x_n > a - \varepsilon$ infinitely often. It follows that $\limsup_n x_n \geq a - \varepsilon > \bar{x}$, contradicting $\limsup_n x_n = \bar{x}$. Hence the assumption that $a > \bar{x}$ leads to contradiction.

It follows that if there is a subsequence $x_{n_k} \rightarrow a$, then $a \leq \limsup_n x_n$, and hence $\limsup_n x_n$ is the biggest number to which $\langle x_n \rangle$ has a convergent subsequence.

—

The following important proposition is left as an exercise.

Proposition A.3.16 A bounded sequence $\langle x_n \rangle_n$ is convergent if and only if $\limsup_n x_n = \liminf_n x_n$.

In that case $\lim_n x_n$ is equal to both $\limsup_n x_n$ and $\liminf_n x_n$.

$$\limsup_n x_n = \lim_n x_n = \liminf_n x_n$$

Exercise A.3.17 Prove Proposition A.3.16.

[Hint: (\Rightarrow): If $\langle x_n \rangle_n$ converges, then every subsequence converges, and to the same limit.

(\Leftarrow) Suppose that $\limsup_n x_n = \liminf_n x_n = x$, and let $\varepsilon > 0$. Then $x_n < x + \varepsilon$ eventually, and $x_n > x - \varepsilon$ eventually.]

□

Hidden away in the analysis of \limsup and \liminf is the following **important** theorem:

Theorem A.3.18 (Bolzano–Weierstrass) Every bounded sequence of reals has a convergent subsequence.

Proof: By Theorem A.3.14, any sequence $\langle x_n \rangle$ has a monotone subsequence. If $\langle x_n \rangle$ is bounded, then so is the subsequence. The result follows by Theorem A.2.4.

◄

A.4 Cauchy Sequences and Completeness

We have already seen that any bounded increasing sequence converges (Theorem A.2.4). This allowed us, in Example A.2.6, to conclude that the sequence $\langle (1 + \frac{1}{n})^n \rangle_n$ converges, though we could not see where it converges to. The Completeness Axiom guarantees the existence of a limit, even if we do not know what that limit is.

Like a bounded increasing sequence, a *Cauchy sequence* is a sequence that “ought to” converge. And, as we shall see, a Cauchy sequence *does* converge: The existence of a limit is guaranteed by the Completeness Axiom, even if we do not know what that limit actually is.

Intuitively, a sequence $\langle x_n \rangle_n$ in \mathbb{R} is a Cauchy sequence if its terms lie *eventually arbitrarily close* to each other. This means that from some point onwards, any two terms are “close”. If all terms lie closer and closer together, there should be some point that they are all clustering around, and that point should be the limit of the sequence $\langle x_n \rangle_n$.

All this “ought” and “should” needs to be made precise.

Definition A.4.1 A sequence $\langle x_n \rangle_n$ in \mathbb{R} is called a *Cauchy sequence* if and only if

$$\sup\{|x_n - x_m| : m, n > N\} \rightarrow 0 \quad \text{as } N \rightarrow \infty$$

i.e. iff for every $\varepsilon > 0$ there is an $N \in \mathbb{N}$ such that

$$|x_n - x_m| < \varepsilon \quad \text{whenever } n, m > N$$

i.e. $\forall \varepsilon > 0 \exists N \in \mathbb{N} \forall n \in \mathbb{N} \forall m \in \mathbb{N} [n > N \wedge m > N \rightarrow |x_n - x_m| < \varepsilon]$

Note that *all* terms from some point onwards need to be within ε of each other, not just successive terms. Thus, for example, if $N = 100$, then $|x_{301} - x_{156734}| < \varepsilon$.

Example A.4.2 The sequence $\langle 1 + (-1)^n 2^{-n} \rangle_n$ is Cauchy. Indeed, given $\varepsilon > 0$, we may choose $N \in \mathbb{N}$ such that $2^{-N} < \frac{\varepsilon}{2}$. If $n, m > N$, then (by the triangle inequality)

$$|(1 + (-1)^n 2^{-n}) - (1 + (-1)^m 2^{-m})| \leq 2^{-n} + 2^{-m} \leq 2^{-N} + 2^{-N} < \varepsilon$$

□

Lemma A.4.3 *If $\langle x_n \rangle_n$ is a convergent sequence in \mathbb{R} , then it is a Cauchy sequence.*

Proof: Suppose that $x_n \rightarrow x$, and that we are given $\varepsilon > 0$. We must find N such that $|x_n - x_m| < \varepsilon$ whenever $n, m > N$.

Now because $x_n \rightarrow x$ there is $N \in \mathbb{N}$ such that $|x_n - x| < \frac{\varepsilon}{2}$ whenever $n > N$. In particular, if $n, m > N$, then

$$|x_n - x_m| \leq |x_n - x| + |x - x_m| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2}$$

Hence $\langle x_n \rangle_n$ is a Cauchy sequence.

⊢

So any convergent sequence is a Cauchy sequence. And this is not surprising: If the terms of a sequence $\langle x_n \rangle_n$ are eventually close to some point x (the limit), then those terms must also eventually be close to each other.

More importantly, the converse is true: Any Cauchy sequence in \mathbb{R} is convergent. To prove this, we will need a number of lemmas. We shall prove:

- Every Cauchy sequence is bounded.
- Every bounded sequence has a convergent subsequence.
- If a Cauchy sequence $\langle x_n \rangle_n$ has a convergent subsequence, then $\langle x_n \rangle_n$ is itself convergent.

Actually, the second point has already been proved. It is the Bolzano–Weierstrass theorem (Theorem A.3.18). Thus we need only prove the first and the last point.

Lemma A.4.4 *If $\langle x_n \rangle_n$ is a Cauchy sequence in \mathbb{R} , then $\langle x_n \rangle_n$ is bounded.*

Proof: Choose $N \in \mathbb{N}$ such that $|x_n - x_m| < 1$ whenever $n, m \geq N$. (This is possible, because $\langle x_n \rangle_n$ is Cauchy — we have taken $\varepsilon = 1$.) Now define

$$K = \max\{|x_1|, |x_2|, \dots, |x_N| + 1\}$$

We show that K is a bound for $\langle x_n \rangle_n$, i.e. that $|x_n| \leq K$ for all $n \in \mathbb{N}$.

Consider separately the two case (i) $n < N$, and (ii) $n \geq N$. In case (i), we obviously have $|x_n| \leq K$, by definition of K . Suppose therefore, that $n \geq N$. In that case, both n and N are $\geq N$, and thus

$$|x_n| \leq |x_n - x_N| + |x_N| \leq 1 + |x_N| \leq K$$

which finishes case (ii).

⊢

Lemma A.4.5 *If $\langle x_n \rangle_n$ is a Cauchy sequence, and if $\langle x_n \rangle_n$ has a convergent subsequence, then $\langle x_n \rangle_n$ itself converges.*

Proof: Suppose that $\langle x_{n_k} \rangle_k$ is a subsequence of the Cauchy sequence $\langle x_n \rangle_n$, and that $x_{n_k} \rightarrow x$ (as $k \rightarrow \infty$). We show that $x_n \rightarrow x$ (as $n \rightarrow \infty$).

So let $\varepsilon > 0$. We must show that there is $N \in \mathbb{N}$ such that $|x_n - x| < \varepsilon$ whenever $n > N$. Now because $\langle x_n \rangle_n$ is a Cauchy sequence, we can find an N_1 such that

$$n, m > N_1 \quad \text{implies} \quad |x_n - x_m| < \frac{\varepsilon}{2}$$

Because $x_{n_k} \rightarrow x$, we can find a K such that

$$k > K \quad \text{implies} \quad |x_{n_k} - x| < \frac{\varepsilon}{2}$$

Now define $N = \max\{N_1, n_K\}$, and let $n > N$. Choose $k > K$ large enough so that also $n_k > N$. Then (i) $n, n_k > N_1$, and (ii) $k > K$. It follows that

$$|x_n - x| \leq |x_n - x_{n_k}| + |x_{n_k} - x| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2}$$

whenever $n > N$.

—

Theorem A.4.6 *Let $\langle x_n \rangle_n$ be a sequence in \mathbb{R} . Then $\langle x_n \rangle_n$ converges if and only if it is a Cauchy sequence.*

Proof: (\Rightarrow) is Lemma A.4.3.

(\Leftarrow) : If $\langle x_n \rangle_n$ is a Cauchy sequence, then it is bounded (by Lemma A.4.4). Hence it has a convergent subsequence (by Theorem A.3.18). It follows that $\langle x_n \rangle_n$ converges (by Lemma A.4.5).

—

The fact that Cauchy sequence converge in \mathbb{R} is depends very much on the Completeness Axiom. If you look back over the proof of Theorem A.4.6, you will not see the Completeness Axiom mentioned explicitly. But we *do* use the Bolzano–Weierstrass Theorem, whose proof requires the Completeness Axiom. This is made clear by the following exercise:

Exercise A.4.7 We define temporarily the following notions for subsets $A \subseteq \mathbb{R}$. We say that

- $A \subseteq \mathbb{R}$ is a *complete set* if and only if whenever $B \subseteq A$ is bounded above (below) then $\sup B \in A$ ($\inf B \in A$).
- We say that $A \subseteq \mathbb{R}$ is a *Cauchy set* if and only if whenever $\langle a_n \rangle_n$ is a Cauchy sequence in A , then $\langle a_n \rangle_n$ converges to an element a in A .

(1.) Which of the following sets are complete; which are Cauchy?

$$[0, 1], \quad (0, 1), \quad \mathbb{R}, \quad \mathbb{Q}, \quad \left\{ \frac{1}{n} : n \in \mathbb{N} \right\}, \quad \{0\} \cup \left\{ \frac{1}{n} : n \in \mathbb{N} \right\}$$

(2.) We prove that $A \subseteq \mathbb{R}$ is a Cauchy set iff it is a complete set.

- (a) Suppose that $A \subseteq \mathbb{R}$ is a complete set, and let $\langle a_n \rangle_n$ be a Cauchy sequence in A . Explain why $\langle a_n \rangle_n$ has a monotone subsequence $\langle a_{n_k} \rangle_k$. Explain why $\lim_k a_{n_k} \in A$. Conclude that $\lim_n a_n \in A$.

- (b) Suppose that $A \subseteq \mathbb{R}$ is a Cauchy set, and that $B \subseteq A$ is bounded above. Explain why there is a sequence $\langle a_n \rangle_n$ in A such that $a_n \rightarrow \sup B$. Conclude that $\sup B \in A$.

□

Exercises A.4.8 1. (a) Prove that if $\langle x_n \rangle$ converges, then $\lim_n (x_{n+1} - x_n) = 0$.

(b) Does the converse hold? i.e., does $\lim_n (x_n - x_{n+1}) = 0$ imply that $\langle x_n \rangle$ converges?

2. (a) Suppose that a sequence $\langle x_n \rangle$ has $|x_{n+1} - x_n| \leq 2^{-n}$ for all $n \in \mathbb{N}$. Show that $\langle x_n \rangle$ converges.

(b) Does the same hold if we only know that $|x_n - x_{n+1}| \leq \frac{1}{n}$ for all $n \in \mathbb{N}$?

□

Appendix B

Sets and Logic

B.1 Logic, Formal Languages, Quantifiers

A *formal language* is a collection of \mathcal{L} whose *logical symbols* include

- **Logical Connectives**

\wedge	and
\vee	or
\rightarrow	implies
\leftrightarrow	if and only if
\neg	not

It is enough to use just two connectives, e.g. \wedge and \neg . We can then define the remainder by

$$\begin{aligned}\varphi \vee \psi &\equiv \neg(\neg\varphi \wedge \neg\psi) \\ \varphi \rightarrow \psi &\equiv \neg\varphi \vee \psi \\ \varphi \leftrightarrow \psi &\equiv (\varphi \rightarrow \psi) \wedge (\psi \rightarrow \varphi)\end{aligned}$$

Just a reminder: \vee is *inclusive-or*: $p \vee q$ is true if and only if at least one of p, q is true, possibly both.

- **Quantifiers**

\forall	For all
\exists	There exists

We have

$$\forall x\varphi \equiv \neg\exists x(\neg\varphi) \quad \exists x\varphi \equiv \neg\forall x(\neg\varphi)$$

- **Variables**

$x, y, z, x_1, x_2, x_3 \dots$

- **Identity relation**

A special binary relation symbol denoted $=$.

Logical symbols have the same meaning, regardless of context. \mathcal{L} also has *non-logical* symbols, whose meaning depends on context:

- **Relation symbols**

For example, if we want to talk about partial orderings, we will want a symbol \leq ; if we want to talk about sets, we will want symbols \in and \subseteq .

- **Function symbols**

For example, if we want to talk about arithmetic, we will want binary function symbols $+$, \times . We may want unary function symbols $-$, $^{-1}$. If we want to talk about sets, we will want binary function symbols \cap , \cup , unary function symbols c , \mathcal{P} ;

- **Constant symbols**

These are specially named elements, and are often regarded as *nullary* function symbols. For example, if we want to talk about addition, a *distinguished element* denoted by 0 plays an important role. If we want to talk about sets, the set \emptyset deserves its own name.

A formal language will generally not contain all of the above non-logical symbols, only those needed to talk about the domain of discourse. \mathcal{L} will also have brackets $(,), [,]$, etc.

The symbols of a formal language may be “strung” together to form two types: *terms* and *formulas*.

- **Terms** are defined as follows:

- (i) Every variable and every constant is a term;
- (ii) If t_1, \dots, t_n are terms, and if F is an n -ary function symbol, then $F(t_1, \dots, t_n)$ is a term;
- (iii) A string is a term only if it can be shown to be so by a finite number of applications of (i) and (ii);

- **Formulas** are defined as follows:

- (i) If t_1, \dots, t_n are terms, and if R is an n -ary relation symbol, then $R(t_1, \dots, t_n)$ is a formula. (This includes the case where R is the logical binary relation symbol $=$).
- (ii) If φ, ψ are formulas, then so are $(\varphi \wedge \psi)$, $(\varphi \vee \psi)$, $(\varphi \rightarrow \psi)$, $(\varphi \leftrightarrow \psi)$;
- (iii) If φ is a formula, then so is $\neg\varphi$;
- (iv) If φ is a formula and x is a variable, then $\forall x\varphi$ and $\exists x\varphi$ are formulas;
- (v) A string is a formula only if it can be shown to be so by a finite number of applications of (i)-(iv).

We often omit brackets when there is no danger of confusion. Moreover, we may also abbreviate $\forall x\forall y\varphi$ by $\forall x, y\varphi$.

If φ is a formula, we write $\varphi(x, y, z)$ to show that the variables of φ are (amongst) x, y, z .

Example B.1.1 Partial orderings

Consider the following language \mathcal{L} : In addition to the logical symbols, \mathcal{L} has a single binary relation symbol \leq . There are no function and constant symbols. Thus the only terms of \mathcal{L} are the variables. Some example of formulas are

$$x \leq y, \quad \forall x(x \leq y \wedge y \leq z) \rightarrow \exists z(\neg(z \leq x))$$

The theory of partial orderings has the following axioms

- (i) $\forall x(x \leq x)$;
- (ii) $\forall x, y(x \leq y \wedge y \leq x \rightarrow x = y)$;

(iii) $\forall x, y, z (x \leq y \wedge y \leq z \rightarrow x \leq z)$.

This theory has many *interpretations*. One is the two-element chain $C_2 = \{0, 1\}$ with $0 \leq 1$. This is a linear ordering, i.e. it satisfies the axiom $\forall x, y (x \leq y \vee y \leq x)$. Another example is the powerset $\mathcal{P}(A)$ of a set A , where \leq is interpreted as “subset”. This ordering is non-linear if A has more than one element.

Thus different structures may satisfy the same axioms.

□

Finally, a note about negating quantifiers: A negation sign can “creep” past a quantifier, but it *flips* the quantifier in the process:

$$\neg \forall x \varphi \equiv \exists x (\neg \varphi) \quad \neg \exists x \varphi \equiv \forall x (\neg \varphi)$$

For example,

$$\begin{aligned} \neg [\forall x \exists y (y > x)] &\equiv \exists x \neg [\exists y (y > x)] \\ &\equiv \exists x \forall y (y \not> x) \end{aligned}$$

B.2 Basic Set Theory

B.2.1 Sets

In the early twentieth century, the following principle was established:

All mathematical objects are sets.

All mathematical notions can be expressed as relationships between sets.

Intuitively, a set is just a collection of objects.

If A is a set and x is some mathematical object, then we say

$$x \in A \quad (x \text{ is an } \textit{element} \text{ of } A)$$

if x is amongst the objects that are collected in A .

A set is characterized entirely by its elements. Two sets are the same if and only if they have the same elements:

$$A = B \iff \forall x [x \in A \leftrightarrow x \in B]$$

Instead of *set*, we will also say *collection*, *family* or *class*. Instead of saying *x is an element of A* , we may also say *x is a member of A* or *x belongs to A* .

We say that a set A is a *subset* of a set B if and only if every element of A belongs to B

$$A \subseteq B \iff \forall x [x \in A \rightarrow x \in B]$$

Thus

$$A = B \text{ iff } (A \subseteq B) \wedge (B \subseteq A)$$

We say that A is a *proper subset* of B if $A \subseteq B$, but $A \neq B$. We also write $B \supseteq A$ to mean $A \subseteq B$.

There are two ways to represent a set:

- (i) By listing its elements: $A = \{a_i : i \in I\}$

(ii) By some defining *property*: $A = \{x : \phi(x)\}$

The following sets have special symbols associated with them:

- The *empty set* $\emptyset = \{\} = \{x : x \neq x\}$.
- $\mathbb{N} := \{1, 2, 3, \dots\}$ is the set of natural numbers.
- $\mathbb{Z}^+ := \{0, 1, 2, 3, \dots\}$ is the set of non-negative integers.
- $\mathbb{Z} := \{\dots, -2, -1, 0, 1, 2, \dots\}$ is the set of integers.
- $\mathbb{Q} := \{\frac{p}{q} : p \in \mathbb{Z}, q \in \mathbb{N}\}$ is the set of rational numbers.
- \mathbb{R} is the set of real numbers.
- \mathbb{C} is the set of complex numbers.

B.2.2 Union and intersection

The symbols \cup, \cap denote, respectively, the *union* and *intersection* of two sets:

$$\begin{aligned} A \cup B &= \{x : x \in A \vee x \in B\} \\ A \cap B &= \{x : x \in A \wedge x \in B\} \end{aligned}$$

The symbols \bigcup and \bigcap denote, respectively, the *union* and *intersection* of a *family* of sets:
If $\mathcal{A} = \{A_i : i \in I\}$ is a family of sets, then

$$\begin{aligned} \bigcup \mathcal{A} &= \bigcup_{i \in I} A_i = \{x : \exists i \in I : x \in A_i\} \\ \bigcap \mathcal{A} &= \bigcap_{i \in I} A_i = \{x : \forall i \in I : x \in A_i\} \end{aligned}$$

Thus

$$\begin{aligned} x \in \bigcup \mathcal{A} &\Leftrightarrow \exists A \in \mathcal{A} [x \in A] \\ x \in \bigcap \mathcal{A} &\Leftrightarrow \forall A \in \mathcal{A} [x \in A] \end{aligned}$$

Note that

$$A \cup B \quad \text{instead of} \quad \bigcup \{A, B\}$$

etc.

We may also write

$$\bigcup_{n=1}^{\infty} A_n = A_1 \cup A_2 \cup A_3 \cup \dots$$

for $\bigcup_{n \in \mathbb{N}} A_n$, etc.

B.2.3 Set difference, complementation and symmetric difference

If A, B are sets, we define the *set difference* of A and B by

$$A - B = \{x : x \in A \wedge x \notin B\}$$

We define the *symmetric difference* of A, B by

$$A \Delta B = (A - B) \cup (B - A) = (A \cup B) - (A \cap B)$$

Often, we will be working with subsets of some *universal set* Ω . If $A \subseteq \Omega$, we define the *complement* of A by

$$A^c = \Omega - A$$

Note that if $A, B \subseteq \Omega$, then

$$A - B = A \cap B^c$$

B.2.4 Set algebra

Note the following *laws*:

- Idempotent laws:

$$A \cup A = A \quad A \cap A = A$$

- Commutative laws:

$$A \cup B = B \cup A \quad A \cap B = B \cap A \quad A \Delta B = B \Delta A$$

- Associative laws:

$$A \cup (B \cup C) = (A \cup B) \cup C \quad A \cap (B \cap C) = (A \cap B) \cap C \quad A \Delta (B \Delta C) = (A \Delta B) \Delta C$$

- Distributive laws:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \quad A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

$$A \cap \bigcup_{i \in I} B_i = \bigcup_{i \in I} (A \cap B_i) \quad A \cup \bigcap_{i \in I} B_i = \bigcap_{i \in I} (A \cup B_i)$$

$$(A \Delta B) \cap C = (A \cap C) \Delta (B \cap C) \quad (A \Delta B) \cup C = (A \cup C) \Delta (B \cup C)$$

- Absorption laws:

$$A \cap (A \cup B) = A \quad A \cup (A \cap B) = A$$

- Complementation laws:

$$A \cap A^c = \emptyset \quad A \cup A^c = \Omega \text{ (the universal set)}$$

$$(A^c)^c = A \quad (A \Delta B)^c = A^c \Delta B$$

- De Morgan's laws:

$$(A \cap B)^c = A^c \cup B^c \quad (A \cup B)^c = A^c \cap B^c$$

$$\left(\bigcup_{i \in I} A_i \right)^c = \bigcap_{i \in I} A_i^c \quad \left(\bigcap_{i \in I} A_i \right)^c = \bigcup_{i \in I} A_i^c$$

- Set difference laws

$$A - (B \cup C) = (A - B) \cap (A - C) \quad A - (B \cap C) = (A - B) \cup (A - C)$$

$$A - \bigcup_{i \in I} B_i = \bigcap_{i \in I} (A - B_i) \quad A - \bigcap_{i \in I} B_i = \bigcup_{i \in I} (A - B_i)$$

- Symmetric difference laws:

$$A \Delta A = \emptyset \quad A \Delta \emptyset = A \quad A \Delta \Omega = A^c$$

Note also that the trivial fact that $A \Delta B \subseteq A \cup B$ implies the following useful *triangle inequality*:

$$A \Delta C \subseteq (A \Delta B) \cup (B \Delta C)$$

Indeed, we have

$$A\Delta C = (A\Delta\emptyset)\Delta B = A\Delta(B\Delta B)\Delta C = (A\Delta B)\Delta(B\Delta C) \subseteq (A\Delta B) \cup (B\Delta C)$$

B.2.5 Products

Since a set is determined completely by its elements, we see that $\{a, b\} = \{b, a\}$ — the order of a, b doesn't matter. It is often convenient to have a structure in which order *does* matter, however. An *ordered pair* (a, b) should be thought of as a collection containing a and b , *in that order*. Thus $(a, b) \neq (b, a)$.

Typically, the an ordered pair (a, b) is defined as follows:

$$(a, b) = \{\{a\}, \{a, b\}\}$$

This is done purely to maintain the principle that all mathematical objects must be sets. The exact definition of an ordered pair doesn't matter, however. What is important is that

$$(a, b) = (c, d) \iff a = c \wedge b = d$$

We will agree to let $(a, (b, c))$ and $((a, b), c)$ denote the same object: In each case, we have a followed by b followed by c . It is convenient to omit the inner pair of brackets, and to denote this object by the ordered triple (a, b, c) .

In the same way, an ordered n -tuple (a_1, a_2, \dots, a_n) should be thought of as a collection containing a_1, a_2, \dots, a_n , *in that order*. We may even consider an I -indexed ordered tuple $(a_i)_{i \in I}$: This is just like the I -indexed set $\{a_i : i \in I\}$, except that the order matters. Note that if the index set I is the set of natural numbers, then $(a_n)_{n \in \mathbb{N}}$ is just a *sequence*:

$$(a_n)_{n \in \mathbb{N}} = a_1, a_2, a_3, \dots$$

Given two set A, B , we define their *cartesian product* by

$$A \times B = \{(a, b) : a \in A \wedge b \in B\}$$

In the same way, we can define the cartesian product of a family of sets:

$$\prod_{i \in I} A_i = \{(a_i)_{i \in I} : a_i \in A_i \text{ for all } i \in I\}$$

Note that an element of $\prod_{i \in I} A_i$ can be thought of as a function: The ordered tuple $(a_i)_{i \in I}$ corresponds to that function $f : I \rightarrow \bigcup_{i \in I} A_i$ with the property that

$$f(i) = a_i$$

Such an f is called a *choice function*. Thus $\prod_I A_i$ is precisely the set of all choice functions $I \rightarrow \bigcup_{i \in I} A_i$.

We also define *powers* of sets as follows:

$$A^I := \prod_{i \in I} A_i \quad \text{where } A_i := A \quad \text{for all } i \in I$$

Note that an element of A^I can be thought of as a function: The ordered tuple $(a_i)_{i \in I}$ corresponds to that function $f : I \rightarrow A$ with the property that

$$f(i) = a_i$$

Thus A^I is precisely the set of all functions from I to A .

B.3 The Extended Real Number System

The extended real line

$$\bar{\mathbb{R}} = [-\infty, \infty] = \mathbb{R} \cup \{\infty, -\infty\}$$

This can be topologized as follows: The family of all sets of the forms

$$[-\infty, a) \quad (a, b) \quad (b, \infty] \quad a, b \in \mathbb{R}$$

forms a base for the topology.

We can also define partial arithmetic operations, as follows: $+$, \cdot agree with ordinary addition and multiplication when applied to real numbers. Also

$$a \pm \infty \quad \infty \cdot a = \infty + a = \pm\infty \quad a \in \mathbb{R}$$

$$a \cdot \infty = \begin{cases} +\infty & \text{if } a > 0 \\ 0 & \text{if } a = 0 \\ -\infty & \text{if } a < 0 \end{cases}$$

Analogous results hold for multiplication with $-\infty$.

Similarly

$$\infty + \infty = \infty \cdot \infty = \infty \quad \infty \cdot -\infty = -\infty \cdot \infty = -\infty$$

The expression

$$\infty - \infty$$

is left undefined.

It is straightforward to check that the commutative, associative, and distributive laws hold in $\bar{\mathbb{R}}$ whenever both sides of the identity under consideration are defined.

Note that we have defined $0 \cdot \infty = 0$.